A Compartmental Model for Computing Cell Numbers in CFSE-based Lymphocyte Proliferation Assays

H.T. Banks and W. Clayton Thompson

Center for Research in Scientific Computation and Center for Quantitative Sciences in Biomedicine

North Carolina State University, Raleigh, NC 27695-8212

Cristina Peligero, Sandra Giest, Jordi Argilaguet and Andreas Meyerhans

ICREA Infection Biology Lab, Department of Experimental and Health Sciences

Univ. Pompeu Fabra, 08003 Barcelona, Spain

January 31, 2012

Abstract

Some key features of a mathematical description of an immune response are an estimate of the number of responding cells and the manner in which those cells divide, differentiate, and die. The intracellular dye CFSE is a powerful experimental tool for the analysis of a population of dividing cells, and numerous mathematical treatments have been aimed at using CFSE data to describe an immune response [29, 30, 31, 36, 37, 40, 46, 47]. Recently, partial differential equation structured population models, with measured CFSE fluorescence intensity as the structure variable, have been shown to accurately fit histogram data obtained from CFSE flow cytometry experiments [18, 19, 50, 52]. In this report, the population of cells is mathematically organized into compartments, with all cells in a single compartment having undergone the same number of divisions. A system of structured partial differential equations is derived which can be fit directly to CFSE histogram data. From such a model, cell counts (in terms of the number of divisions undergone) can be directly computed and thus key biological parameters such as population doubling time and precursor viability can be determined. Mathematical aspects of this compartmental model are discussed, and the model is fit to a data set. As in [18, 19], we find temporal and division dependence in the rates of proliferation and death to be essential features of a structured population model for CFSE data. Variability in cellular autofluorescence is found to play a significant role in the data, as well. Finally, the compartmental model is compared to previous work, and statistical aspects of the experimental data are discussed.

Key words: Cell proliferation, cell division number, CFSE, label structured population dynamics, partial differential equations, inverse problems.

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
31 JAN 2012		**00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
A Compartmental Model for Computing Cell Numbers in CFSE-based Lymphocyte Proliferation Assays	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
North Carolina State University,Center for Research in Scientific Computation,Department of Mathematics,Raleigh,NC,27695-8212	**CRSC-TR12-03**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Some key features of a mathematical description of an immune response are an estimate of the number of responding cells and the manner in which those cells divide, differentiate, and die. The intracellular dye CFSE is a powerful experimental tool for the analysis of a population of dividing cells, and numerous mathematical treatments have been aimed at using CFSE data to describe an immune response [29, 30, 31, 36, 37, 40, 46 47]. Recently, partial differential equation structured population models, with measured CFSE fluorescence intensity as the structure variable, have been shown to accurately fit histogram data obtained from CFSE flow cytometry experiments [18, 19, 50, 52]. In this report, the population of cells is mathematically organized into compartments, with all cells in a single compartment having undergone the same number of divisions. A system of structured partial differential equations is derived which can be fit directly to CFSE histogram data. From such a model, cell counts (in terms of the number of divisions undergone) can be directly computed and thus key biological parameters such as population doubling time and precursor viability can be determined. Mathematical aspects of this compartmental model are discussed, and the model is fit to a data set. As in [18, 19], we find temporal and division dependence in the rates of proliferation and death to be essential features of a structured population model for CFSE data. Variability in cellular autofluorescence is found to play a significant role in the data, as well. Finally, the compartmental model is compared to previous work and statistical aspects of the experimental data are discussed.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**42**	

1 Introduction

The human immune response is a complex process in which the behavior of individual cells in the lymphatic system is altered by a multitude of intra- and extracellular signals. The mathematical analysis of lymphocyte activation and division can be performed on a wide range of scales, from the molecular level (antigen presentation and recognition) to the population level. This report focuses on the latter. To that end, one can look at the total number of divisions a cell has undergone since activation and how cells in different generations differ in phenotype.

The development by Lyons and Parish [53] of the intracellular dye carboxyfluorescein succinimidyl ester (CFSE) for use in proliferation assays has resulted in an essential experimental tool for researchers studying these complex processes. CFSE is nonradioactive and provides long-lasting, bright, and relatively uniform labeling of all cells in a population without adversely affecting the internal machinery of the cells. When cells divide, the dye is partitioned approximately in half. Thus, when labeled cells are stimulated to divide, the CFSE content of individual cells in the population can be assessed via flow cytometry and the number of divisions a cell has undergone can be determined by comparing the measured fluorescence intensity of a cell to the measured fluorescence intensity of an undivided cell [54, 53, 59, 60, 70, 72]. When individual cell fluorescence intensity measurements for all cells in a given population are binned into a histogram, each generation of cells appears as a "peak" in the histogram data. The data set used in this report is the same as that from [18, 66] and the experimental protocol is discussed at length there. The data is depicted in Figure 1.

While the quantitative modeling of CFSE data has traditionally focused on the deconvolution of the data into numbers of cells per generation [29, 30, 31, 36, 37], recent efforts [18, 19, 50, 52] have used a structured population model in order to fit the CFSE histogram data directly. Using this technique, we have produced a strong, physically and biologically motivated model which is quite capable of replicating the observed CFSE histogram data obtained via flow cytometry. The most recent partial differential equation (PDE) model is a fragmentation equation which relates the structured population density $n(t, x)$ to the rates of proliferation $\alpha(t, x)$

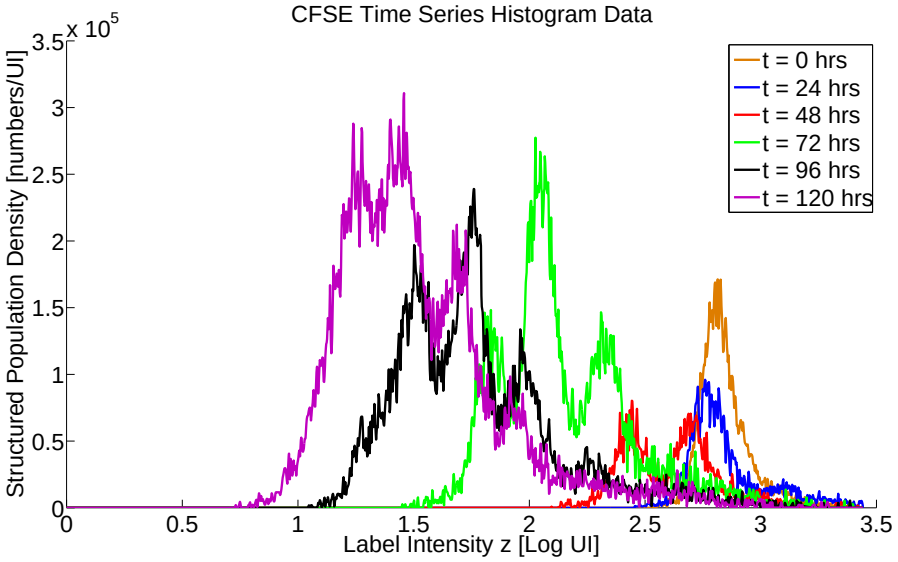

Figure 1: Data set for a CFSE-based proliferation assay. Note that the data is presented in the logarithmic coordinate $z = \log_{10}(x)$, in units log UI. As cells divide, CFSE is diluted and the initially unimodal population density becomes multimodal. While it is easy to distinguish the various peaks in the data, the overlap between peaks results in some systematic error when attempting to identify a region of the horizontal axis with a specific division number. In addition to this overlap, the slow drift to the left over time (as a result of intracellular turnover of CFSE) further weakens the correlation.

and death $\beta(t,x)$ under the assumption of Gompertz decay of label and is given by

$$\frac{\partial n(t,x)}{\partial t} - ce^{-kt}\frac{\partial[(x-x_a)n(t,x)]}{\partial x} = -(\alpha(t,x)+\beta(t,x))n(t,x) + \chi_{[x_a,x^*]}4\alpha(t,2x-x_a)n(t,2x-x_a). \quad (1)$$

The structure variable x is the fluorescence intensity (in arbitrary units of intensity, UI) of the cells. Because this fluorescence intensity arises primarily from CFSE within the cell, we refer to this as a label structured population model (as opposed to age or size structure, etc. [56]). It is known that cells lose FI in time even in the absence of division as a result of the natural decay of CFSE and the turnover of intracellular proteins to which the fluorescent conjugates bind. The advection term in the equation above accounts for this phenomenon using a Gompertz [38] decay velocity $v(t,x) = -c(x-x_a)e^{-kt}$ with characteristic parameters c and k, which has been shown [18] to accurately describe the biphasic decay [55, 59, 70] of CFSE FI observed in data sets. The parameter x_a represents the natural autofluorescence intensity of cells in the absence of CFSE, assumed for the moment in (1) to be constant across the cell population.

The goal of such a mathematical model is to provide biologists with simple yet intuitive and meaningful parameters with which a population of dividing cells can be described. In particular, information such as average rates of division and cell viability are essential to the analysis of the effects of changing experimental conditions (e.g., differences in donors, differences between diseased and healthy cells) on proliferative behavior. The motivation for the use of FI as a structure variable is that the serial dilution of CFSE by cell division creates a correlation between measured FI and the number of divisions a cell has undergone. Thus the proliferation and death rate functions $\alpha(t,x)$ and $\beta(t,x)$, which are estimated in terms of the structure variable x as well as time, can be used to compute average division rates in terms of the number of divisions undergone [18]. (For instance, if $x > 1000$ UI corresponds to undivided cells, then the average proliferation rate, as a function of time, for undivided cells is the average value of $\alpha(t,x)$ in the region $x > 1000$.)

This motivating assumption is accurate to a degree, as one can clearly discern the distinct generations of cells in the data set depicted in Figure 1. However, the peaks corresponding to particular generations of cells overlap slightly and drift to the left in time (as a result of CFSE decay), thus weakening the correlation between the state variable and division number. In [7, 18], it is shown that the proliferation and death rates can be parameterized with respect to a 'translated variable' which accounts for the loss of measured FI in time, and that this translated variable is more strongly correlated with division number than the original structure variable x. Yet, the overlap between distinct peaks in the data remains problematic, and it is not clear how much error may be introduced into the estimated proliferation and death rates by this overlap of distinct generations.

Furthermore, while the model (1) is advantageous in being able to estimate average proliferation and death rates without any deconvolution of the data into cell numbers, it cannot be used to accurately assess the number of cells in a particular generation. This information could be approximated by integrating the structured density $n(t,x)$ over a region $[x_1, x_2]$ (corresponding approximately to the location of a given peak in the histogram data), but this approximation is limited by the extent to which distinct generations of cells in the histogram data overlap. Traditional deconvolution techniques (such as fitting peaks with normal or lognormal curves) impose particular forms on the experimental data which may bias the computed number of cells in each generation.

While all these efforts to date correspond to several iterations in an iterative modeling process (for a philosophical discussion see [20, Chapter 1]) to attempt to understand cell proliferation using CFSE labeling of populations, we clearly have not yet reached a satisfactory understanding of the complex phenomena involved. The fragmentation models used with the CFSE data can be considered as what have been termed *Aggregate Data/Aggregate Dynamics* or Type II inverse problems as presented in [1, Chapter 14] and [5]. Such problems are also common in investigations with models for electromagnetic propagation in inhomogeneous dielectric materials including biotissue [13, 14], vibrational dissipation in viscoelastic materials [16], and HIV cellular progression models [1, 4, 5]. To better understand rates at the generation number cohort or division number cohort level, one should attempt to develop individual (cohort) dynamics to investigate the CFSE data in a Type I framework of *Aggregate Data/Individual (Cohort) Dynamics* inverse problems such as those discussed in [1, Chapter 14] and [5]. Similar approaches have been successfully pursued in marine and insect population models [3, 6, 10, 12, 21] as well as in physiologically based pharmacokinetic (PBPK) models in toxicology [5, 17]. Fortunately, a simple reformulation of (1) allows such an approach and permits both the accurate quantification of total cells per division number and the accurate estimation of proliferation and death rates in terms of division number in such a framework.

Rather than modeling the *population* with a single differential equation, one can model each individual *generation* of cells with a single equation,

$$\frac{\partial n_i}{\partial t} + \frac{\partial [v(t,x) n_i(t,x)]}{\partial x} = -(\alpha_i(t) + \beta_i(t)) n_i(t,x) + R_i(t,x),$$

with the generations linked through the division mechanism $R_i(t,x)$ as a source term (see the next section). This is a common technique in existing ordinary and delay differential equations models for dividing cells (see [25, 29, 30]). Because each generation of cells is assigned to a particular compartment (indexed by i) with unique proliferation and death rates, it is not necessary to estimate these rates in terms of the structure variable x, so that peak overlap and label decay no longer affect the accuracy of the estimated rates. This is in contrast to previous work [18, 19] in which considerable space is devoted to answering the question of how to parameterize the structural dependence of the proliferation and death rates. As an added advantage, the number of parameters necessary for the parameterizations of the proliferation and death rates is reduced (because there is no longer a need to parameterize the functions α_i and β_i in terms of the structure variable). Furthermore, the existence of multiple compartments makes it possible to accurately determine cell numbers in terms of divisions undergone, even though the computed densities (for the distinct compartments) will still overlap when placed simultaneously on the x axis. Because this model does not rely upon any assumptions as to the shape (normal, lognormal, etc.) of the generation peaks (instead starting from an initial condition and fitting directly to the CFSE histogram data) systematic bias should be avoided.

In this report, we begin with a careful formulation of the compartmental model. The solution to this model is then presented and computational aspects are discussed. Next we establish an inverse problem for the estimation of the AutoFI and Gompertz parameters, as well as the proliferation and death rate functions $\alpha_i(t)$ and $\beta_i(t)$. As in previous work [18, 19], multiple parameterizations of the proliferation and death rate functions are considered with the goal of determining how these rates depend on both division number and on time. After presenting results which demonstrate the enhanced capabilities of the compartmental model, the statistical properties of the flow cytometry data are considered and ramifications for the quantification of uncertainty in the estimated parameters are discussed.

2 The Compartmental Model

The derivation of the compartmental model follows immediately from the derivation of the fragmentation model (1) in [18], which is itself a variation of the structured population models of Bell-Anderson [22] and Sinko-Streifer [63]. A complete derivation of the compartmental model can be found in Chapter 3 of [66]. Let $n_i(t,x)$, $0 \leq i \leq i_{\max}$ be the label structured population density of a population of cells stained with CFSE and having undergone i divisions. The structure variable x is the fluorescence intensity (FI) of a cell (in arbitrary units of intensity, UI) satisfying $x \geq x_a$ where x_a is the natural autofluorescence intensity (AutoFI) of cells; t is time (in hours). While it is known that AutoFI increases significantly when cells become activated, this increase is not believed to be significant for the current modeling effort (as AutoFI contributes minimally to the measured FI of a labeled but unactivated cell). Thus, the parameter x_a should be understood to describe AutoFI for *activated* cells. It is known that FI scales linearly with the concentration of CFSE used to label a population of cells, and that this measurement does not change significantly when cells increase in size [54]. Thus we assume FI is a mass-like quantity.

The label-structured density of a population of dividing cells is modeled by the system of PDEs

$$\begin{aligned}
\frac{\partial n_0}{\partial t} + \frac{\partial [v(t,x) n_0(t,x)]}{\partial x} &= -(\alpha_0(t) + \beta_0(t)) n_0(t,x) \\
\frac{\partial n_1}{\partial t} + \frac{\partial [v(t,x) n_1(t,x)]}{\partial x} &= -(\alpha_1(t) + \beta_1(t)) n_1(t,x) + R_1(t,x) \\
&\vdots \\
\frac{\partial n_{i_{\max}}}{\partial t} + \frac{\partial [v(t,x) n_{i_{\max}}(t,x)]}{\partial x} &= -\beta_{i_{\max}}(t) n_{i_{\max}}(t,x) + R_{i_{\max}}(t,x)
\end{aligned} \tag{2}$$

where $v(t,x)$ is the natural rate of CFSE FI decay (as a result of the turnover of CFSE within individual cells) and $R_i(t,x) = 4\alpha_{i-1}(t)n_{i-1}(t,2x-x_a)$ for $1 \leq i \leq i_{\max}$ represents the influx of newly divided cells. Note the assumption that $\alpha_{i_{\max}} = 0$. While there is no mathematical limit to the number of generations which can be computed, experimental data generally exhibits fewer than 10 divisions. In an inverse problem setting (see Section 3), the parameter $i_{\max}$ can be easily fixed in advance by simply counting the number of generations which appear in the data. Because it is then known that there are no cells with generation number $i_{\max}+1$, there must be no proliferation in generation $i_{\max}$, and the model can be simplified by setting $\alpha_{i_{\max}} = 0$. Of course, the process of determining $i_{\max}$ could be automated via model refinement statistical tests, but that seems unnecessary given the ease with which the parameter can be identified from data.

There is an additional mathematical justification for setting $\alpha_{i_{\max}} = 0$. The total quantity of CFSE FI in the population is

$$M(t) = \int_{x_a}^{\infty} x \left(\sum_{i=0}^{i_{\max}} n_i(t,x) \right) dx.$$

Using the definition of $M(t)$ and the system of equations (2), we can show that

$$\begin{aligned} \frac{dM}{dt} = &\int_{x_a}^{\infty} v(t,x) \left(\sum_{i=0}^{i_{\max}} n_i(t,x) \right) - \int_{x_a}^{\infty} x \left(\sum_{i=0}^{i_{\max}} \beta_i(t) n_i(t,x) \right) \\ &+ x_a \int_{x_a}^{\infty} \left(\sum_{i=0}^{i_{\max}} \alpha_i(t) n_i(t,x) \right) - \int_{x_a}^{\infty} x \left(\alpha_{i_{\max}}(t) n_{i_{\max}}(t,x) \right). \end{aligned}$$

While the first three terms on the right side of this equation are physically relevant and expected (loss of FI by Gompertz decay, loss of FI by death, and the additive role of AutoFI, respectively) the final term is not experimentally valid because cells do not recognize a maximum division number after which they must leave the measured population. The requirement that $\alpha_{i_{\max}} = 0$ eliminates this term.

The initial condition must be prescribed for each i,

$$n_i(0,x) = \Phi_i(x). \tag{3}$$

It will generally (but not necessarily always) be true that $\Phi_i(x) = 0$ for $i \geq 1$ (that is, all cells in the population are undivided at $t = 0$). These initial condition curves are determined from data taken at $t = 0$ (see Section 2.2). The left ($x = x_a$) boundary conditions are the no-flux boundary conditions

$$v(t,x_a) n_i(t,x_a) = 0 \tag{4}$$

for all $0 \leq i \leq i_{\max}$. Because the problem is defined on the semi-infinite domain $x \geq x_a$, these conditions are sufficient to compute a solution. This is in contrast to previous work [18, 19, 50, 52] in which a zero-recruitment boundary condition, $n(t,x_{\max}) = 0$, is imposed at the right boundary of the computational domain. As discussed in the next section, under appropriate conditions these two formulations are equivalent.

2.1 Model Solution

For many decay velocities $v(t,x)$ of interest, the system of equations (2) can be solved analytically using the method of characteristics. As discussed previously, we assume that the rate at which cells naturally lose FI is described by the same function, $v(t,x)$, for all cells independent of division number. As such, the characteristic lines are the same for each generation of cells. Furthermore, it will be assumed that this rate of FI loss is adequately described by a Gompertz decay process [38]; this has been shown [18] to effectively describe the biphasic decay [55, 59, 70] characteristic of proliferation assay data when the intracellular label is CFSE. Thus we have

$$v(t,x) = -c(x - x_a)e^{-kt} \tag{5}$$

where $c > 0$ and $k > 0$ (both with units 1/hr) are parameters to be estimated using the data. In effect, this function describes cellular FI which decreases exponentially (with initial rate c) to the level of cellular AutoFI,

while the exponential rate itself decreases (exponentially) with rate k. The assumption of Gompertz decay of cellular FI has the additional benefit of trivially satisfying the left boundary condition (4) for all i, provided $n_i(t, x_a)$ is finite (so that the flux at the boundary is well-defined).

Incorporating the Gompertz decay process, we can rewrite the system (2) as

$$\begin{aligned}
\frac{\partial n_0}{\partial t} - ce^{-kt}(x - x_a)\frac{\partial n_0}{\partial x} &= -(\alpha_0(t) + \beta_0(t) - ce^{-kt})n_0(t, x) \\
\frac{\partial n_1}{\partial t} - ce^{-kt}(x - x_a)\frac{\partial n_1}{\partial x} &= -(\alpha_1(t) + \beta_1(t) - ce^{-kt})n_1(t, x) + R_1(t, x) \\
&\vdots \\
\frac{\partial n_{i_{\max}}}{\partial t} - ce^{-kt}(x - x_a)\frac{\partial n_{i_{\max}}}{\partial x} &= -(\beta_{i_{\max}}(t) - ce^{-kt})n_{i_{\max}}(t, x) + R_{i_{\max}}(t, x).
\end{aligned} \tag{6}$$

The characteristic lines (for all i) are described by

$$\frac{dx}{dt} = v(t, x) = -c(x - x_a)e^{-kt}, \tag{7}$$

and hence the characteristic line emanating from the point $(0, s)$ in the tx-plane is

$$x(t; s) = x_a + (s - x_a)\exp\left[-\frac{c}{k}\left(1 - e^{-kt}\right)\right], \tag{8}$$

where $s \geq x_a$ parameterizes the line along which the initial condition is prescribed.

Define

$$f_i(t) = \alpha_i(t) + \beta_i(t) - ce^{-kt}.$$

For undivided cells ($i = 0$), the solution along a characteristic line emanating from a point $(0, s)$ in the tx-plane is given by

$$\frac{\partial n_0}{\partial t} = -f_0(t)n_0(t, x(t; s))$$

with $n_0(0, x(0; s)) = n_0(0, s) = \Phi_0(s)$. Thus the solution along characteristic lines is

$$n_0(t, x(t; s)) = \Phi_0(s)\exp\left(-\int_0^t f_0(\tau)d\tau\right). \tag{9}$$

As written above, the system of equations (6) is defined on the semi-infinite domain $x \geq x_a$. In general, the initial condition function $\Phi_0(x)$, can be determined from data (see Section 3) only on some finite segment $[x_a, x_{\max}]$ of the domain. However, there is no loss of generality in extending the initial condition curve by assuming $\Phi_0(x) = 0$ if $x > x_{\max}$. This is in contrast to [18, 19, 50, 52] in which a PDE was defined only on the finite interval $[x_a, x_{\max}]$ and a zero-recruitment boundary was imposed. In fact, the two formulations are equivalent provided $\Phi(x_{\max}) = 0$ (in the former models; $\Phi_i(x_{\max}) = 0$ for all i in the compartmental model) and $v(t, x) < 0$. As the semi-infinite formulation is notationally simpler and easy to implement, we use it here.

The solutions for $i \geq 1$ along the same characteristic lines (8) are described by

$$\frac{\partial n_i}{\partial t} = -f_i(t)n_i(t, x(t; s)) + R_i(t, x(t; s)) \tag{10}$$

with $n_i(0, x(0; s)) = \Phi_i(s)$ and the solutions are

$$n_i(t, x(t; s)) = \Phi_i(s)\exp\left(-\int_0^t f_i(\tau)d\tau\right) + \int_0^t R_i(\tau, x(\tau; s))\exp\left(-\int_\tau^t f_i(\xi)d\xi\right)d\tau. \tag{11}$$

It is worth noting that the solution by the method of characteristics involves the construction of an integral surface in the coordinates t and s. The change of coordinates from t and x to t and s has Jacobian

$$J = \begin{vmatrix} \frac{\partial t}{\partial t} & \frac{\partial t}{\partial s} \\ \frac{\partial x}{\partial t} & \frac{\partial x}{\partial s} \end{vmatrix} = \exp\left(-\frac{c}{k}\left(1 - e^{-kt}\right)\right),$$

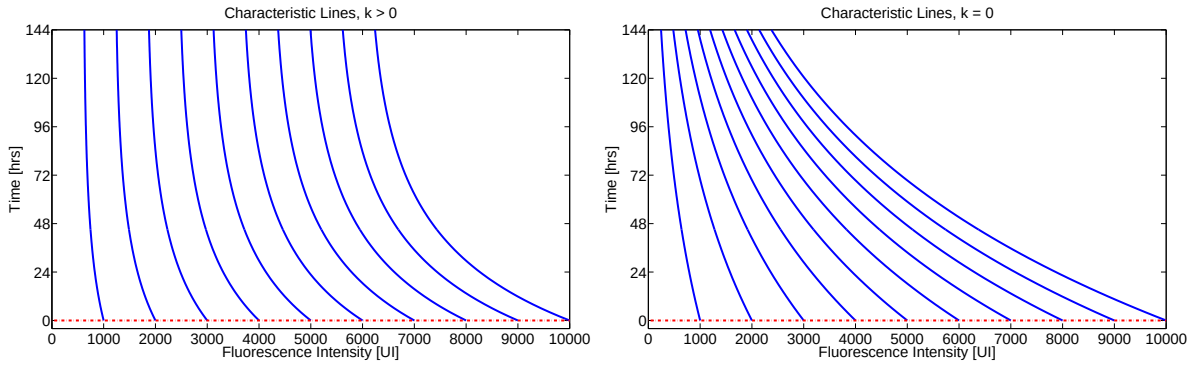

Figure 2: Characteristic lines given by Equations (7)-(8) when $c = 1 \times 10^{-2}$ and $k = 2 \times 10^{-2}$ (left) and in the limiting case when $k = 0$ (right). Notice that distinct characteristic lines will remain separated by some positive distance for all time in the former case, while in the latter case the lines asymptotically converge to $x = x_a$. The horizontal broken line along the bottom of both graphics is the line along which the initial condition data is given. It is clear that this initial condition curve is nowhere tangent to a characteristic line, hence the local existence of a unique solution is guaranteed.

which is nonsingular along the initial condition curve ($t = 0$). Hence we are guaranteed (by the construction above) that a unique solution exists at least locally near the initial condition curve. Note that in the limit as $k \to 0^+$, the Jacobian is $J_{k\downarrow 0} = e^{-ct}$, which becomes singular asymptotically in time (reflecting the asymptotic convergence of the characteristic lines). In such a case, one might observe solutions which grow without bound. This is only of minimal concern, however, as the total label loss resulting from decay is small over the duration of a typical experiment. Possible characteristic lines (for $k > 0$ and $k = 0$) are shown graphically in Figure 2.

For the remainder of this document, it will be assumed that all cells are undivided at $t = 0$, so that $\Phi_i(x) = 0$ for $i \geq 1$. This condition is satisfied by essentially all experimental data. Thus, the only nontrivial initial condition for the PDE system (6) is $\Phi_0(x)$. As this model is motivated by an attempt to fit and explain experimental data, this smooth initial condition must be constructed from data taken at the beginning of the experiment. Our process for doing so is described below, followed by the numerical algorithm for computing the solutions (9) and (11).

2.2 Initial Condition Construction

For the construction of the initial condition, we use experimental data (which is noisy histogram data in the logarithmic coordinate $z = \log_{10}(x)$) collected at $t = 0$ hours in order to determine $\Phi_0(x)$. The data consist of ordered pairs (z_k^0, n_k^0), which denote the number of cells n_k^0 counted into the histogram bin (subject to measurement error) with its left boundary at z_k^0 when $t = 0$. In order to obtain a smooth initial condition function from the noisy data (z_k^0, n_k^0), a smooth line is drawn through the original histogram data which is taken to represent the 'true' cell counts in the absence of noise. The numerical values are recovered from the smooth line using DataThief [67] to form the 'noiseless' counts $(z_k^0, \hat{n}_k^0)$, which are then easily transferred from the logarithmic coordinate resulting in new ordered pairs $(x_k^0, \hat{n}_k^0)$ (because the $\hat{n}_k^0$ are approximate numbers of counted cells as opposed to a structured density, the values do not need to be rescaled when changing from z to x).

Finally, we must use these 'noiseless' cell counts in the x coordinate in order to determine the structured density initial condition $\Phi_0(x)$ for (9). To do so, we first define the function

$$\varphi(x) = \sum_k \hat{n}_k^0 l_k(x),$$

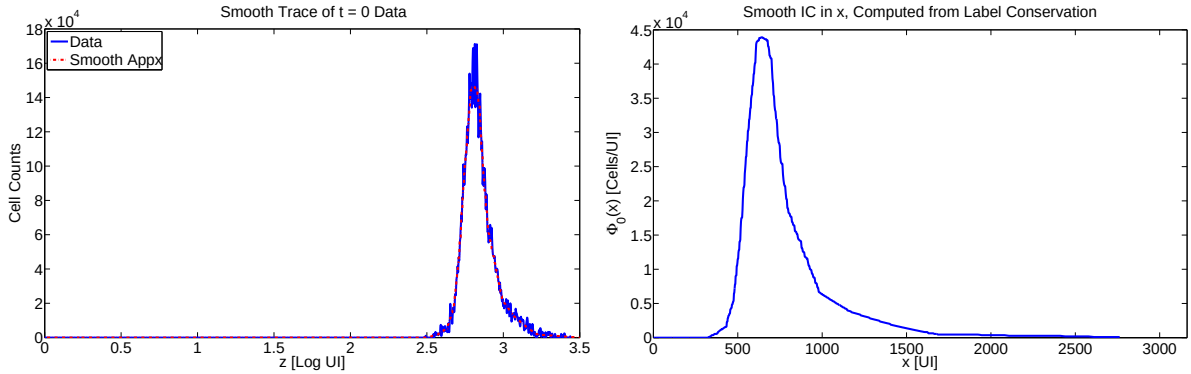

Figure 3: Left: Smooth curve drawn through the experimental data taken at $t = 0$ hours. Right: Initial condition function $\Phi_0(x)$ computed from the smooth line using the algorithm of Section 2.2.

where $l_k(x)$ are piecewise linear functions satisfying

$$l_j(x_k^0) = \begin{cases} 1, & j = k \\ 0, & j \neq k \end{cases}.$$

Thus the function $\varphi(x)$ is a piecewise linear function such that $\varphi(x_k^0) = \hat{n}_k^0$. Next, we compute the total measured FI in the population at $t = 0$ using the original noisy data,

$$FI_{data} = \sum_k x_k^0 n_k^0.$$

Similarly, the total FI in the smooth data function $\varphi(x)$ is

$$FI_{smooth} = \int x\varphi(x)dx,$$

where the integral is approximated using the composite trapezoidal rule. The initial condition function is then constructed as

$$\Phi_0(x) = \frac{FI_{data}}{FI_{smooth}}\varphi(x).$$

The results of this technique are shown in Figure 3. This method ensures that the total measured FI in the initial condition curve is equal to the total measured FI in the original noisy data. Because the mathematical model (2) is derived from conservation principles (considering FI as a mass-like quantity), this provides a useful comparison between the data and the model, as well as a method to assess the accuracy of the numerical simulations. It is worth noting that such a complex procedure is unnecessary in the event the histogram bins are evenly spaced (in the logarithmic coordinate z). In such an event, a smooth density function (in z) can be computed from the smooth cell counts $(z_k^0, \hat{n}_k^0)$ simply by dividing the counts by the bin spacing. The function $\Phi_0(x)$ can then be computed as a simple change of variables from z to x. However, this method of computation may result in discontinuous jumps in the computed density if there are abrupt changes in the sizes of adjacent histogram bins. Moreover, total measured FI for the initial condition curve will not necessarily be equal to the total FI in the noisy data (although the two values should still be close) if such a method were to be used. Thus, we find the rescaling method above to be preferable even in the case of evenly spaced bins.

2.3 Numerical Solution

Given the initial condition function $\Phi_0(x)$ as computed above, it now remains to numerically compute the solutions (9) and (11) for $n_0(t, x)$ and $n_i(t, x)$, $1 \leq i \leq i_{\max}$, respectively. In the structure variable, the solutions are

computed on a fixed (i.e., one that does not change with division number) mesh $\{x^{(k)}\}$, $1 \leq k \leq N_x$ (these should be distinguished from the $x_k^j = 10^{z_k^j}$ used to describe the data). While it is not strictly necessary for each compartment to be computed on the same grid, there seems to be little advantage in varying the structure variable mesh with division number. Because the major features (the 'peaks') of the structured density solution shrink by approximately a factor of two with each division, it is advantageous to choose the points $\{x^{(k)}\}$ so that they are logarithmically spaced (that is, so that the collection $\{\log_{10}(x^{(k)})\}$ is evenly spaced). This ensures an increasing density of nodes as x decreases, and hence as the major features of the solution become more condensed. The uneven spacing of the nodes for the structural variable does not cause any numerical difficulties as the algorithm presented below requires only interpolation (i.e., no finite-difference derivatives) in the structural dimension. Because the model presented in this report uses distinct compartments for each generation of cells, and because each generation of cells remains in a relatively small region in the structure variable (see [18, 19, 50, 52], which were motivated by this fact), it is certainly true that the use of a single structural mesh for all compartments results in some unnecessary storage and computations. However, the value of N_x has only a small effect on computational time (see below) so that this is of little concern.

In time, the solutions are computed on a fixed, evenly spaced mesh $\{t^{(m)}\}$, with spacing h_t. Unlike time-stepping finite difference methods (such as the Lax-Wendroff method used in [18, 19, 52] which require storage of the solution at only the most recent time steps, the method of characteristics solution (11) requires an integration along characteristic lines over the history of the solution (see Equation (13) below). This integration is computed via the trapezoidal rule, using quadrature nodes which correspond to the time mesh $\{t^{(m)}\}$. While this method of computing the solution is storage-intensive, the requirements are not unreasonable, even when running in MATLAB on a 32-bit desktop machine.

It is obvious from Equations (9) and (11) that the system (6) can be solved inductively on i. For time $t^{(m)}$ and FI $x^{(k)}$, we find from Equation (8)

$$s(t^{(m)}, x^{(k)}) = (x^{(k)} - x_a)\exp\left(\frac{c}{k}\left(1 - e^{-kt^{(m)}}\right)\right) + x_a. \tag{12}$$

The solution $n_0(t, x)$ (Equation (9)) is then computed by multiplying the values of $\Phi_0(s)$ by the scalar

$$\exp\left(-\int_0^{t^{(m)}} f_0(\tau)d\tau\right) = \exp\left(-\int_0^{t^{(m)}} \left(\alpha_0(\tau) + \beta_0(\tau) - ce^{-k\tau}\right) d\tau\right).$$

For all parameterizations of the functions $\alpha_i(t)$ and $\beta_i(t)$ considered in this report, the integral above can be computed exactly.

As noted above, it is assumed $\Phi_i(x) = 0$ for $i \geq 1$. Thus the solutions $n_i(t, x)$ can be rewritten

$$n_i(t, x) = \int_0^t G_i(\tau; t, x)d\tau \tag{13}$$

where

$$G_i(\tau; t, x) = 4\alpha_{i-1}(\tau)n_{i-1}(\tau, 2x(\tau, s) - x_a)\exp\left(-\int_\tau^t f_i(\xi)d\xi\right).$$

As above, the scalar $4\alpha_{i-1}(\tau)\exp\left(-\int_\tau^t f_i(\xi)d\xi\right)$ is computed exactly. The values of the function $n_{i-1}(t, x)$, though already computed, will only be available at discrete points $(t^{(m)}, x^{(k)})$ and thus (13) must be computed via quadrature. Because every solution $n_i(t, x)$ is computed on the same, evenly spaced time mesh, a simple solution is to use this same time mesh (with the trapezoidal rule) in order to approximate the integral. Thus, given a point $(t^{(l)}, x^{(k)})$, $l \leq m$, we must first determine s according to (12). This is then used to compute

$$\tilde{x}^{(l)} = 2x(t^{(l)}, s) - x_a,$$

for $0 \leq l \leq m$, where $x(t, s)$ is given in Equation (8). Thus we have

$$G_i(t^{(l)}; t^{(m)}, x^{(k)}) = 4\alpha_{i-1}(t^{(l)})n_{i-1}(t^{(l)}, \tilde{x}^{(l)})\exp\left(-\int_{t^{(l)}}^{t^{(m)}} f_i(\xi)d\xi\right),$$

with the values of $n_{i-1}(t^{(l)}, \tilde{x}^{(l)})$ determined by linear interpolation. Finally, from (13),

$$n_i(t,x) = \int_0^t G_i(\tau;t)d\tau \approx$$
$$= h_t \cdot \sum_{l=1}^{m-1} G_i(t^{(l)};t^{(m)},x^{(k)}) + \frac{h_t}{2}\left(G_i(t^{(0)};t^{(m)},x^{(k)}) + G_i(t^{(m)};t^{(m)},x^{(k)})\right). \quad (14)$$

We now consider how the computed solution changes as the mesh parameters N_x and h_t are changed. As a test case, nominal parameters (for x_a, c, k, $\alpha_i(t)$ and $\beta_i(t)$) were used to compute a solution at $t = 120$ hours using various combinations of values for N_x and h_t. The results are shown in comparison in Figure 4. For convenience, the solutions $n_i(t,x)$ have been summed together and graphed in terms of the log FI ($z = \log_{10}(x)$) coordinate.

As noted above, the algorithm does not require any quadrature or finite differences in the structural component. At most, it is necessary to use interpolation (linear interpolation seems sufficient) in order to compute the function G_i above in the likely event $\tilde{x}^{(l)} \notin \{x^{(k)}\}$. Thus, one would expect approximately second-order accuracy in N_x. In fact, we find (computationally) this expectation is exceeded. The explanation lies in the iterative manner in which the solution is computed. Consider computing $n_1(t,x)$ (Equation (13)), provided $n_0(t,x)$ is already computed. On one hand, because this computation will require (linear) interpolation, we would expect the resulting error to depend upon the mesh-spacing of $n_0(t,x)$. However, $n_0(t,x)$ is computed from the initial condition function $\Phi_0(x)$ which is defined (see Section 2.2) as a piecewise linear function. It follows that, as N_x approaches the number of points used in defining $\Phi_0(x)$, the error will no longer decrease (because a piecewise linear function is being used to approximate a more coarsely defined piecewise linear function). In this report, the function $\Phi_0(x)$ is defined with 806 points. Thus, it is no surprise that we find little difference in the solutions computed with $N_x = 512$ and $N_x = 1024$.

The use of the trapezoidal rule with step size h_t to approximate the integral in (13) results in a numerical solution which is second order in h_t. Also, we see that, at each time step $t^{(m)}$ ($0 \le m \le T/h_t$, for a solution computed on $t \in [0,T]$), Equation (14) requires m computations of the function $G_i(\tau;t,x)$. Thus we expect the computational time to scale as $O(1/h_t^2)$. Table 1 summarizes the average computational time for various combinations of N_x and h_t. As expected, computational time approximately quadruples as h_t is halved. As the algorithm in the previous section has been fully vectorized, N_x has only a minimal effect on computational time.

When fitting the compartmental model to data in an inverse problem setting, we must balance the need for an accurate solution with the desire to quickly evaluate the model (given a set of parameters). As such, in the results presented in Section 4, we use $N_x = 512$ with $h_t = 0.5$ hours.

3 Inverse Problem Formulation

We now consider the inverse problem of calibrating the model (6) to a particular data set. As stated previously, the data consist of ordered pairs (z_k^j, n_k^j) indicating the total number of cells n_k^j counted into the histogram bins with left boundary at z_k^j (in the log FI coordinate) at time t_j. The notation is meant to emphasize the possibility that the histogram bins need not share a common fixed width, nor need they be the same at each measurement time. The data set we will use to calibrate the compartmental model is shown in Figure 5, with measurements taken at $t = 24$, 48, 96, and 120 hours.

Let $n_i(t,x)$ be the solution of the compartmental model for cells having undergone i divisions. Then the total population of cells is

$$n(t,x) = \sum_{i=0}^{i_{\max}} n_i(t,x).$$

Because this model solution is computed in the linear FI coordinate x while the data is given in the logarithmic FI coordinate $z = \log_{10}(x)$, we define

$$\tilde{n}(t,z) = 10^z \ln(10) n(t,x(z)) = 10^z \ln(10) n(t,10^z). \quad (15)$$

The function $\tilde{n}(t,z)$ is the structured population density in terms of the new structure variable z. The factor $10^z \ln(10)$ arises from the chain rule in the integral formulation of the model (see Section 2) and is needed to

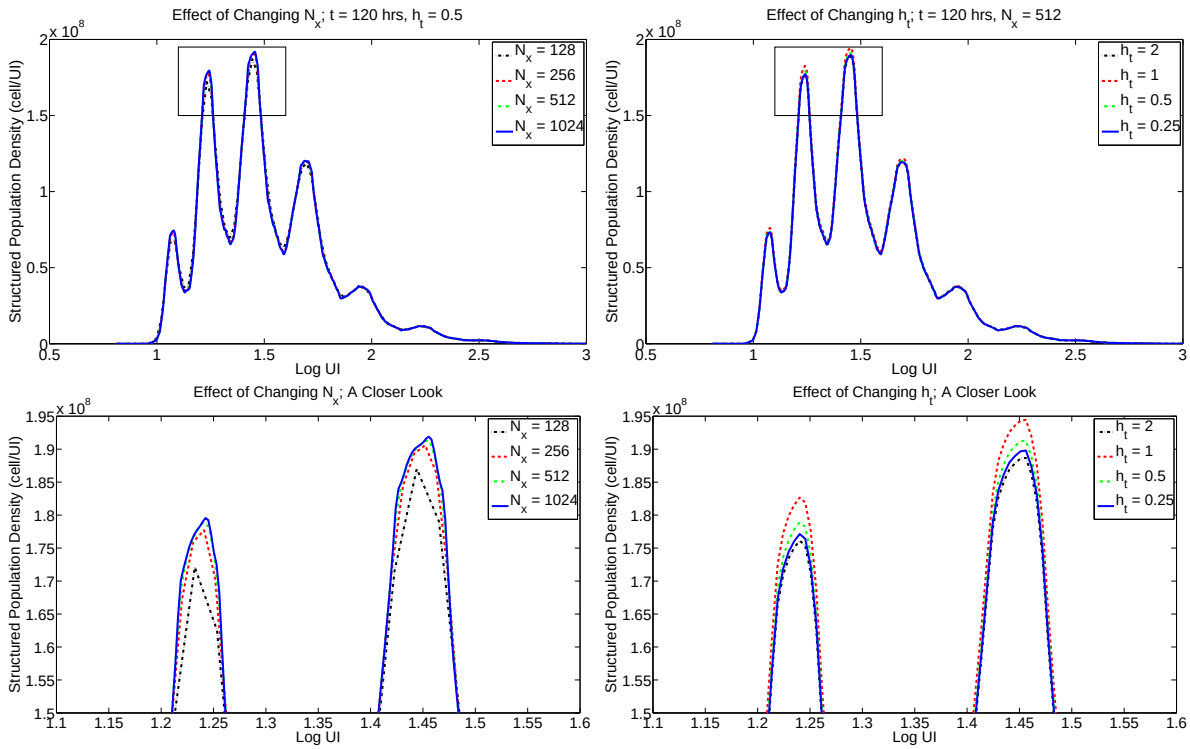

Figure 4: Left: Effect of changing the number of structure variable nodes N_x with the time increment fixed at $h_t = 0.5$ hours. The computed solutions are similar for all values of N_x shown (top). Zooming in (bottom), we find that there is only a small difference (less than 2% max) in the computed solutions for $N_x \geq 256$ with solutions for $N_x = 512$ and $N_x = 1024$ virtually indistinguishable. Right: Effect of changing the time increment h_t with the number of structural variable nodes fixed at $N_x = 512$. While the difference between the solution computed for $h_t = 2$ and $h_t = 1$ is large, there is a much smaller difference (approximately 1% max) between $h_t = 0.5$ and $h_t = 0.25$, as expected. Note that the proximity of the $h_t = 2$ solutions to the $h_t = 0.25$ solution is mere coincidence and does not hold more generally.

Table 1: Effects of h_t and N_x on computational time. Computational times are shown in seconds, with h_t specified in hours. As expected, computational time is quadratic in h_t. Meanwhile, N_x has a much smaller effect on computational time.

$N_x \setminus h_t$	2.00	1.00	0.50	0.25
128	2.1	7.5	29.8	120.0
256	2.3	8.9	35.6	150.7
512	2.7	10.9	44.8	199.7
1024	3.9	16.8	69.7	297.9

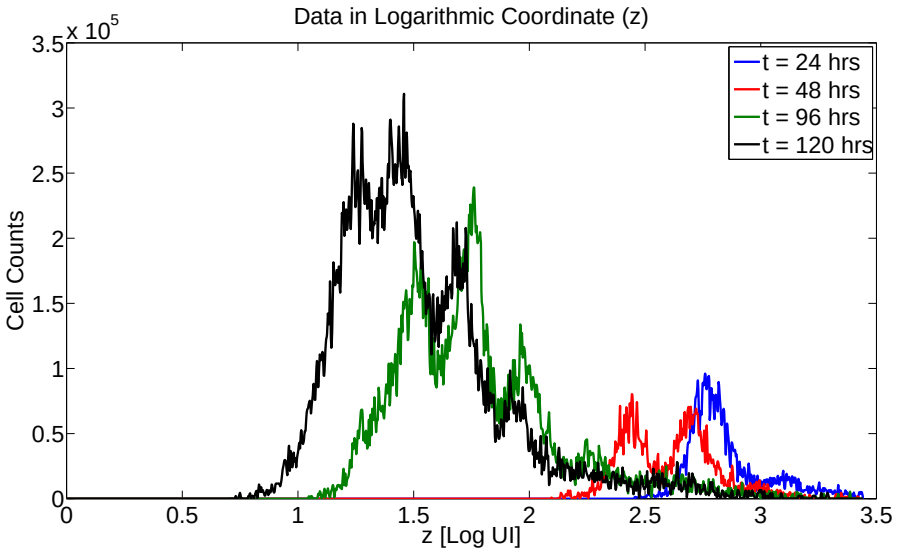

Figure 5: CFSE data set for the compartmental model.

conserve the quantity of label after the change of variables. Finally, we need to convert this structured density into cell counts for comparison with the data. Thus we define

$$I[\tilde{n}](t_j, z_k^j) \equiv \int_{z_k^j}^{z_{k+1}^j} \tilde{n}(t_j, z) dz,$$

which is the observation operator for the compartmental model. In practice, because the transformed model solution $\tilde{n}(t, z)$ is computed only at discrete points (t_j, z_k^j), we must approximate this observation operator,

$$I[\tilde{n}](t_j, z_k^j) \approx I_A[\tilde{n}](t_j, z_k^j) = \left[\frac{\tilde{n}(t_j, z_{k+1}^j) + \tilde{n}(t_j, z_k^j)}{2}\right] \left(z_{k+1}^j - z_k^j\right). \tag{16}$$

3.1 Ordinary Least Squares

Given an initial condition as constructed in Section 2.2, the solution $n(t, x)$ (and hence, $\tilde{n}(t, z)$) is completely determined by the parameters x_a (AutoFI), c and k (Gompertz decay), as well as the proliferation rates $\{\alpha_i(t)\}$ and the death rates $\{\beta_i(t)\}$. Let $\theta = \{x_a, c, k, \{\alpha_i(t)\}, \{\beta_i(t)\}\} \subset \Theta$, where Θ is some set of admissible values for θ. (While it will be necessary to make some simplifying assumptions on Θ in order to render the inverse problem computationally tractable, we postpone that discussion for the moment and proceed with a general overview of the inverse problem procedure.) Thus we may write the model as $n(t, x; \theta)$. The goal of the inverse problem is to determine some value of the parameter θ which minimizes the distance (in an appropriate sense) between the cell counts determined by the model solution, $I[\tilde{n}](t_j, z_k^j)$, and the histogram data. For this report, we choose least squares as the method of estimation. Following standard inverse problem procedure [20, 27, 28, 62], we define the random variables

$$N_k^j = I[\tilde{n}](t_j, z_k^j; \theta_0) + \mathcal{E}_{kj}, \tag{17}$$

where $\mathcal{E}_{kj}$ are independent random variables satisfying $E[\mathcal{E}_{kj}] = 0$ representing measurement error and/or 'noise' in the data. The parameter θ_0 is the 'true' parameter (given the model) which is assumed to exist and to describe the data. The data, then, represent a single realization of these random variables,

$$n_k^j = I[\tilde{n}](t_j, z_k^j; \theta_0) + \epsilon_{kj}.$$

The assumption that the data are generated from the specified model, given a nominal truth parameter, is common in inverse problem formulations [8, 20]. While θ_0 is generally unknown, we can define the estimator

$$\theta_{WLS} = \arg\min_{\theta\in\Theta} \sum_{k,j} \frac{R_{kj}^2}{w_{kj}} = \arg\min_{\theta\in\Theta} \sum_{k,j} \frac{1}{w_{kj}} \big(I[\tilde{n}](t_j, z_k^j; \theta) - N_k^j\big)^2, \tag{18}$$

which minimizes the weighted sum (with weights w_{kj}^{-1}) of squared residuals R_{kj}. Because the N_k^j are random variables, so are the R_{kj} and, hence, so is θ_{WLS}. Using the data, we may obtain the estimate

$$\hat{\theta}_{WLS}(n_k^j) = \arg\min_{\theta\in\Theta} \sum_{k,j} \frac{r_{kj}^2}{w_{kj}} = \arg\min_{\theta\in\Theta} \sum_{k,j} \frac{1}{w_{kj}} \big(I[\tilde{n}](t_j, z_k^j; \theta) - n_k^j\big)^2.$$

In theory, the weights w_{kj}^{-1} should be chosen to reflect the variance of the random variables N_k^j. In fact, the accurate, unbiased estimation of standard errors as well as confidence intervals around parameter estimates is premised upon an accurate statistical model (hence accurate weights) for the error terms $\mathcal{E}_{kj}$. In practice, however, such a statistical model is rarely (if ever) known a priori, and some additional assumptions must be made. For this report, we assume a constant variance (CV) error model, $Var(\mathcal{E}_{kj}) = \sigma_0^2$ for all k and j. In this case, $w_{kj} = 1$ for all k and j and (18) becomes an ordinary least squares (OLS) problem,

$$\theta_{OLS} = \arg\min_{\theta\in\Theta} J(\theta|N_k^j) = \sum_{k,j} \mathcal{R}_{kj}^2 = \arg\min_{\theta\in\Theta} \sum_{k,j} \big(I[\tilde{n}](t_j, z_k^j; \theta) - N_k^j\big)^2, \tag{19}$$

with corresponding estimate

$$\hat{\theta}_{OLS}(n_k^j) = \arg\min_{\theta\in\Theta} J(\theta|n_k^j).$$

The function $J(\theta|n_k^j)$ is the OLS cost of the model, given the data, and is often written simply as $J(\theta)$. The expanded notation is meant to emphasize the dependence of the estimate on the particular data set used to fit the model.

It should be noted that, rather than consider constant variance errors in an OLS framework, one could alternatively consider a statistical model with constant coefficient of variation (CCV), $Var(\mathcal{E}_{kj}) = \sigma_0^2 (I[\tilde{n}](t_j, z_k^j; \theta_0))^2$. Then $w_{kj} = (I[\tilde{n}](t_j, z_k^j; \theta_{GLS}))^2$ and (18) becomes the generalized least squares (GLS) problem defined implicitly by

$$\theta_{GLS} = \arg\min_{\theta\in\Theta} \sum_{k,j} \frac{\mathcal{R}_{kj}^2}{(I[\tilde{n}](t_j, z_k^j; \theta_{GLS}))^2} = \arg\min_{\theta\in\Theta} \sum_{k,j} \frac{\big(I[\tilde{n}](t_j, z_k^j; \theta) - N_k^j\big)^2}{(I[\tilde{n}](t_j, z_k^j; \theta_{GLS}))^2}, \tag{20}$$

with corresponding estimate $\hat{\theta}_{GLS}(n_k^j)$. As noted above, the results presented in Section 4 will focus on parameter estimation in an OLS framework. A more thorough consideration of the reliability of the assumptions for the statistical error model in the inverse problem is postponed until the Discussion. For the moment, we focus on the applicability of the compartmental model to a particular data set–that is, how well the compartmental model fits the data. Of course, the measure of fit is assessed in an OLS framework, which may be slightly different than a GLS or more general WLS framework. The misspecification of the error model is known to result in biased standard errors (and hence inaccurate confidence intervals), and thus no such work is carried out here. (Related efforts on determination of the precise form of measurement error, and hence the corresponding statistical error, in a family of data sets similar to the one used here is being pursued and will be reported on in a separate manuscript.) In spite of this drawback, a slight misspecification of the exact error model should have only minimal effect on the estimated best-fit parameters (see, e.g., the computational example of Section 3.4.2 of [20]), and thus we proceed with the OLS estimation of θ_0.

3.2 Parameterizations of Proliferation and Death Rates

We have already defined the parameter $\theta = \{x_a, c, k, \{\alpha_i(t)\}, \{\beta_i(t)\}\} \subset \Theta$ which describes a given model solution. The parameters x_a, c, and k are all elements of $\mathbb{R}$ (although in a generalization below, we consider estimation of a

probability distribution on the parameter x_a) and thus pose no problem for the estimation procedure. However, the proliferation and death rates $\alpha_i(t)$ and $\beta_i(t)$, $0 \leq i \leq i_{\max}$, are contained in some (infinite-dimensional) function space. Mathematically, the solutions (9) and (11) require only $\alpha_i(t)$, $\beta_i(t) \in L_2(0,T)$ in order for the solution to be well-defined. Because it can reasonably be assumed that these functions are bounded, this condition is naturally met. However, as currently written, (19) contains a minimization over an infinite-dimensional space Θ. In order to make the estimation problem amenable to computation, additional assumptions and/or approximations are necessary.

The primary motivation for using a label-structured PDE model to analyze histogram data from CFSE-based proliferation assays was an attempt to use measured FI as a surrogate for division number and hence to investigate how the proliferation and death rates for a population of cells change with division number. In earlier efforts [18, 19, 50, 52], this was accomplished by allowing the proliferation and death rates to depend explicitly on the state variable (x or z). For the compartmental model formulated in this report, the number of divisions undergone is accounted for directly, so that it is no longer necessary to have the α_i and β_i dependent upon the structure variable (following the assumption that the interference of CFSE with the intracellular machinery is negligible). Additionally, it was found in [18, 19] that explicit time-dependence of the rate of cell proliferation is a significant feature of an accurate label structured PDE model. We would like to pursue this investigation using the new compartmental model. Thus, as we consider possible parameterizations of the proliferation and death rate functions, we do so with an eye toward determining the heterogeneity of the rates (that is, how they vary with division number), as well as the possible time-dependence of the proliferation rates.

We begin with the death rate functions $\beta_i(t)$. It has long been observed that a significant proportion of undivided cells die in the first few days in culture, and that this cell death occurs independent of cellular activation [37]. Beyond these observations, we would like to explore how the death rates of cells depend on division number. Specifically we consider the following possible parameterizations for the death rate functions $\beta_i(t)$:

B1 $\beta_i(t) = 0$ for all i and for all t;

B2 $\beta_i(t) = \beta$ for all i and for all t;

B3 $\beta_0(t) = \beta_0$, $\beta_i(t) = 0$ for $i \geq 1$;

B4 $\beta_0(t) = \beta_0$, $\beta_i(t) = \beta$ for $i \geq 1$;

B5 $\beta_i(t) = \beta_i$ for each i.

The possibility **B1** is included as a baseline for comparison, as a means of concluding the necessity of a death term in the mathematical model. As noted above, it is expected that a model which lacks a mechanism to describe cell death will predict far too many cells in the population when compared to the experimental observations [31, 37]. Parametrization **B2** assumes a constant death rate in the population for all cells regardless of division number. Gett and Hodgkin [37] have shown that parametrization **B3**, in which undivided cells die but all cells which proceed through the first division will remain in the population indefinitely, can be accurately used to predict the number of cells in the population up to approximately 90 hours. More generally, one might consider that cells which have divided at least once may die, but at a rate which is possibly different from the rate for undivided cells. This parametrization **B4** has also been successfully used to model proliferation assay data [29, 30, 31]. Finally, in parametrization **B5**, we consider the possibility that the death rate is completely heterogenous with respect to division number [29, 35, 42, 46].

While the model is derived in sufficiently general terms to include time-dependent death rate functions, we do not consider any such parameterizations in this report. It is certainly possible that, for particular cell lines and under particular culture conditions, feedback mechanisms such as activation-induced cell death may in fact be time-dependent [37]. For a (hypothetical) population of cells which divides almost synchronously, such time-dependence would be identical to division-number-dependence (i.e., a mechanism which does not appear until, say, 90 hours could be equivalently modeled as a mechanism which does not appear until 3 divisions have been completed). Thus it seems reasonable to conclude that, to some extent, the necessity of time-dependent death rates in the mathematical model will depend on the degree of synchronicity observed in the experimental data. At the very least, past experience [18, 19] as well as the results presented here (Section 4) seem to suggest little need for such time-dependence, at least for the current data set.

Table 2: Chosen nodes for the estimation of piecewise linear proliferation rates. Bold font indicates a node for which the proliferation rate was set to zero rather than estimated. For each generation, the proliferation rate is assumed to be zero outside the set of nodes shown. Thus the proliferation rate is estimated at three nodes for each division number.

Generation (i)	$\{t_{\alpha_i}^{(q)}\}$
0	**24**,48,60,72,**96**
1	**48**,60,84,108,**120**
2	**48**,60,84,108,**120**
3	**60**,72,96,120
4	**60**,72,96,120
5	**60**,72,96,120
6	**60**,72,96,120

Unlike for the death rate functions, past experience [18, 19] does indicate a potential need for explicit time-dependence of the proliferation rate functions $\alpha_i(t)$. (The fact that time dependence for cell death rates seems to be sufficiently modeled with only division dependence while a similar result does not hold for the proliferation rates may be explained if the time-dependence of the proliferation rates occurs on a scale faster than the average time a cell takes between subsequent divisions.) To explore possibilities we consider the following possible parameterizations for the proliferation rate functions:

A1 $\alpha_0(t) = \alpha_0$; $\alpha_i(t) = \alpha$ for all i;

A2 $\alpha_i(t) = \alpha_i$ for all t;

A3 $\alpha_0(t) = \alpha_0 \chi_{[t>t^*]}$; $\alpha_i(t) = \alpha$ for all i;

A4 $\alpha_0(t) = \alpha_0 \chi_{[t>t^*]}$; $\alpha_i(t) = \alpha_i$;

A5 piecewise linear functions of time (see below).

Previous authors [31, 37, 40] have emphasized a special importance for the time required for a cell to complete its first division. In case **A1**, it is assumed that undivided cells divide at a rate which may be different than the rate for divided cells, but that neither of these two rates depends on time [30]. Alternatively, we consider the more general case **A2** where each generation of cells divides with its own (time-independent) rate [47]. We also consider a simple time-dependent mechanism in which there is a delay before cells begin to divide. A quick glance at the data (Figure 1) reveals that no division occurs in the population for at least the first 24 hours. Such a delay can be easily incorporated into the model with a step function at some specified time t^*. Previous models [30] have found such a transient in the undivided population to be a significant feature of an accurate mathematical model. The proliferation rates for subsequent generations may **A4** or may not **A3** vary with the number of divisions undergone.

Finally, following the example of [18], we consider using piecewise linear splines to incorporate time-dependence into the proliferation rates. Given a fixed set of nodes $\{t_{\alpha_i}^{(q)}\}$, we consider rates of the form

$$\alpha_i(t) = \sum_q a_i^{(q)} l_i^{(q)}(t),$$

where $l_i^{(q)}(t_{\alpha_i}^{(p)}) = 1$ if $p = q$ and is zero if $p \neq q$. In Table 2 we list the nodes $\{t_{\alpha_i}^{(q)}\}$ used for the estimation of the proliferation rate functions. These particular nodes have been chosen based upon careful consideration of the data in Figure 1 as well as past experience.

Independent of which parameterizations of the proliferation and death rates are used, it should be noted that the current model formulation features proliferation and death rates which are essentially Malthusian in nature (see Section 2). That is, the rates at which cells in a particular generation divide and die is assumed

to be proportional to the total number of cells in that generation (with 'constants' of proportionality $\alpha_i(t)$ and β_i for proliferation and death, respectively). Alternatively, a model can easily be derived with limiting proliferation and death rates (e.g., logistic rates, Gompertz rates, etc.). Malthusian rates have been used with some success in previous models and should be accurate for any population of cells which divides sufficiently rapidly. Biologically speaking, a cell must proceed through several necessary activities (growth, DNA replication, microtubule formation, etc.) between any two divisions, and this must induce some minimum cell cycle time. Tools such as delay differential equations or stochastic processes have been used to mathematically model the cell cycle (see, e.g., [29, 30, 33, 36, 40, 43, 44, 46, 58, 64, 73]) and have resulted in several successful models. We find the current model with its Malthusian rates to be simple and intuitive while also fully capable of accurately fitting the data (see Section 4). However, it is imperative that the parameters estimated when fitting the model to a particular data set be interpreted in the context of the form of the model being used.

3.3 Probabilistically Distributed AutoFI

The derivation and solution of the compartmental model have been given so far under the assumption that the natural brightness of cells in the absence of any CFSE molecules, i.e., the autofluorescence intensity or AutoFI, can be modeled with sufficient accuracy by a single scalar parameter x_a. However, it is known that the AutoFI of a single cell changes as the cell becomes activated, and that AutoFI varies from cell to cell in the population, even among activated cells.

The AutoFI of cells can be measured directly by setting aside a portion of cells from the PBMC culture which are not labeled with CFSE (but which receive an otherwise identical treatment). The results of such a measurement are depicted in Figure 6 for two donors, each at two different measurement times. These data sets were taken independently of the data set shown in Figure 1, which is used to calibrate the model. Because FI measurements are not absolute–they depend on the calibration and gain settings of the flow cytometer at the time of the experiment–the data shown in Figure 6 are intended only to examine the shape of the AutoFI distribution in the population, not its absolute magnitude. As time progresses, the distribution of AutoFI in the data from both donors increases slightly in mean and is increasingly skewed to the right. These features are also found in additional data sets for $24 < t < 144$ (results unpublished) and appear to be the result of some unmodeled biological processes. The most likely explanation is the known increase in AutoFI as cells become activated [54, Fig. 6]. After a sufficient amount of time, essentially all cells in the culture have either become activated or have died.

Following the discussion at the beginning of Section 2, we may consider only AutoFI for activated cells. While we have thus far assumed that this AutoFI can be sufficiently modeled with a single parameter, Figure 6 suggests that we might need to consider a probability distribution on a range of values for the parameter x_a. Let $n(t, x; x_a)$ represent the structured population density of a cohort or subpopulation of cells all of which share the same AutoFI parameter x_a, subject to (6). Assume further that this parameter x_a is distributed in the total population of cells with some probability distribution P. Then it follows that the total population is described by

$$\eta(t, x) = E[n(t, x; x_a)|P] = \int_{x_a^{min}}^{x_a^{max}} n(t, x; x_a) dP(x_a). \tag{21}$$

It is now clear that the structured density $\eta(t, x)$ for the total population of cells will depend upon the probability measure P. Figure 6 depicts the experimental AutoFI data for each donor and measurement time fitted (ordinary least squares) with a scaled lognormal curve. While such an assumption may possibly be of limited validity early in the experiment (probably as a result of the activation process, as discussed above), most cells are undivided at such times and hence the contribution of AutoFI to the total FI of those cells is minimal. Thus we may assume that P is reasonably well-described by a lognormal distribution. Hence

$$\frac{dP}{dx_a} = p(x_a) = \frac{1}{x_a \sigma \sqrt{2\pi}} \exp\left(-\frac{(\log x_a - \mu)^2}{2\sigma^2}\right),$$

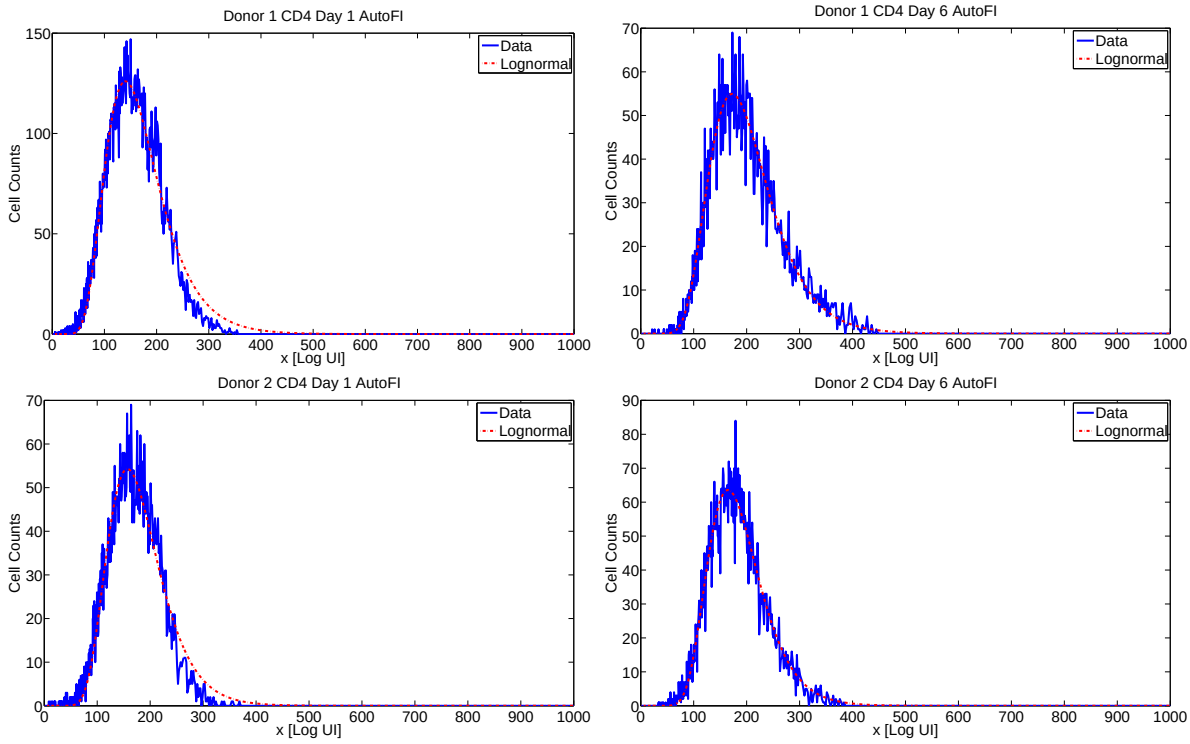

Figure 6: Experimentally determined AutoFI distributions with OLS best-fit scaled lognormal curves. PBMCs of 2 blood donors were cultured without CFSE staining and without stimulation. After 24 (left) and 144 (right) hours respectively, cells were stained for CD4 surface expression and analyzed by flow cytometry. Shown are the histograms of CD4 cell counts as a function of CFSE FI. We see that a lognormal distribution for AutoFI is quite accurate by $t = 144$ hours (right). Such an assumption is less accurate at $t = 24$ hours, when a significant portion of cells in the population remain unactivated.

where

$$\mu = \log(E[x_a]) - \frac{1}{2}\log\left(1 + \frac{Var(x_a)}{E[x_a]^2}\right)$$
$$\sigma^2 = \log\left(1 + \frac{Var(x_a)}{E[x_a]^2}\right).$$

Under such a parametric assumption, the population density $\eta(t,x)$ is uniquely described by the two parameters $E[x_a]$ and $STD[x_a] = \sqrt{Var(x_a)}$ (in addition to the parameters θ discussed so far in this section).

The integral in Equation (21) can be easily computed via the midpoint rule. Let $\{x_a^m\}$ be a set of evenly spaced points with spacing Δx_a. Then

$$\eta(t,x) \approx \sum_{m=1}^{M} n(t,x;x_a^m)p(x_a^m)\Delta x_a. \tag{22}$$

As written, Equation (22) requires the computation of M forward solutions in order to approximate the total population density. However, this computationally intensive approach can be avoided by a change of variables. Define $y = \log_{10}(x - x_a)$ and $\hat{n}(t,y) = 10^y \log(10) n(t,x(y)) = 10^y \log(10) n(t, 10^y + x_a)$. Then the system (6) becomes

$$\begin{aligned}
\frac{\partial \hat{n}_0}{\partial t} - \frac{ce^{-kt}}{\log 10}\frac{\partial \hat{n}_0}{\partial y} &= -(\alpha_0(t) + \beta_0(t) - ce^{-kt})\hat{n}_0(t,x) \\
\frac{\partial \hat{n}_1}{\partial t} - \frac{ce^{-kt}}{\log 10}\frac{\partial \hat{n}_1}{\partial y} &= -(\alpha_1(t) + \beta_1(t) - ce^{-kt})\hat{n}_1(t,x) + 2\alpha_0(t)\hat{n}_0(t, y + \log_{10} 2) \\
&\vdots \\
\frac{\partial \hat{n}_{i_{\max}}}{\partial t} - \frac{ce^{-kt}}{\log 10}\frac{\partial \hat{n}_{i_{\max}}}{\partial y} &= -(\beta_{i_{\max}}(t) - ce^{-kt})\hat{n}_{i_{\max}}(t,x) + 2\alpha_{i_{\max}-1}(t)\hat{n}_{i_{\max}-1}(t, y + \log_{10} 2).
\end{aligned} \tag{23}$$

It is then clearly observed that the parameter x_a no longer appears in the system of equations for the compartmental model in the structure variable y, while only the new initial condition,

$$\hat{\Phi}_0(y) = 10^y \log(10)\Phi_0(10^y + x_a), \tag{24}$$

will now depend on x_a. However, provided the initial uptake of CFSE in the experimental procedure results in cells with measured FI significantly greater than their AutoFI (which is always the case for useful experimental data), $\Phi_0(x) = 0$ unless $x >> x_a$ (and hence, unless $10^y >> x_a$). As such, the dependence of the initial condition on the parameter x_a can be safely ignored. This fact is demonstrated with an example in Figure 7. In general, it is expected that CFSE-labeled cells are approximately 100-1000 times brighter than unlabeled cells (see, e.g., [54, 60, 72]; as mentioned previously, the actual measured FI values depend on machine calibration, and hence will vary from experiment to experiment). Given the initial condition data (Figure 3) for our particular data set of interest, it is reasonable to assume $E[x_a] \sim 10$. In Figure 7, a sample lognormal distribution with $E[x_a] = 12$ and $STD[x_a] = 4$ is depicted on the left. (These values for the mean and standard deviation can be taken as maximum, worst-case bounds. It is expected that the mean value of x_a is no more than 12, with standard deviation less than 4.) We can assess the effect of the parameter x_a on $\hat{\Phi}_0(y)$ by computing $\hat{\Phi}_0(y)$ for extreme values of x_a (that is, values in the far-left and far-right tails of the density function). The resulting functions (as well as a third function, showing $\hat{\Phi}_0(y)$ when $x_a = E[x_a]$) are shown on the right of Figure 7.

It is clear from Figure 7 that the initial condition function (for y as a structure variable) changes only minimally for any reasonable values of x_a. (Moreover, the original initial condition $\Phi_0(x)$ was already approximate, having been computed from data in Section 2.2.) Thus, computationally, when computing the structured population density according to (21), we compute only a single initial condition from Equation (24) using $x_a = E[x_a]$. The system (23) can then be solved to obtain $\hat{n}(t,y)$ (which does not depend on x_a at all). Next, for each value of x_a in (22), one can compute

$$n(t,x;x_a) = \frac{\hat{n}(t,y(x))}{\log(10)(x - x_a)} = \frac{\hat{n}(t,\log_{10}(x - x_a)}{\log(10)(x - x_a)},$$

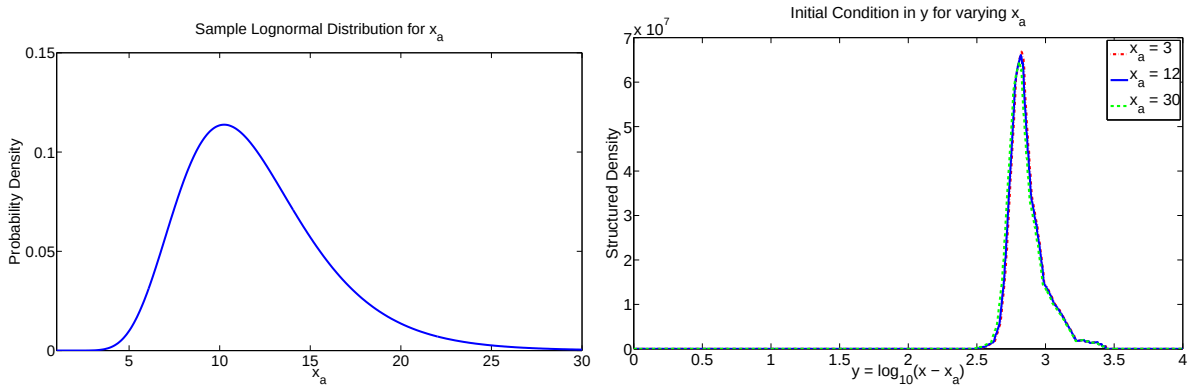

Figure 7: Left: A hypothetical lognormal AutoFI distribution with $E[x_a] = 12$ and $Var(x_a) = 4$. Right: Initial conditions (in the structure variable y) computed for the mean value of x_a (solid line) as well as two for two extreme values of x_a. One can see that the value of the parameter x_a has very little effect on the initial condition $\hat{\Phi}_0(y)$.

in order to determine the population structured density $\eta(t, x)$. It should be noted that, while the change of variables from x to y eliminates the parameter x_a from the system of PDEs, and we have shown that the effect of x_a on the initial condition $\hat{\Phi}_0(y)$ is negligible, it is not true that the parameter x_a can be ignored entirely. The negligible effect of x_a on the initial condition is the result of the brightness of CFSE-labeled cells at the beginning of the experiment. However, as time progresses, CFSE intensity is lost as cells divide and CFSE degrades, so that AutoFI constitutes a larger percentage of the measured FI. In other words, while it is reasonable to assume $n(0, x) = \Phi_0(x) = 0$ unless $x >> x_a$, this assumption does not hold more generally for $n(t, x)$ $(t > 0)$.

Finally, when using Equation (22) to approximate the total population density, one must make certain that the parameter M is sufficiently large to provide desired accuracy. In Figure 8, a sample density is computed at $t = 120$ hours using three different values of M. Given the discussion, above, there is essentially no difference in computational time as M changes. While the solution is not accurately captured for $M = 10$, there is no measurable difference between the solutions for $M = 100$ and $M = 1000$. Henceforth, if it is assumed that AutoFI is distributed in the population of cells, the total population $\eta(t, x)$ will be computed via Equation (22) with $M = 100$.

3.4 Remarks on the Inverse Problem

At this point, we have considered numerous different parameterizations for the proliferation rate functions $\alpha_i(t)$ (**A1**-**A5**), and the death rates β_i (**B1**-**B5**). Each of these parameterizations results in a distinct set of parameters which will need to be estimated from the data. We also have the additional label loss parameters c and k, as well as the AutoFI parameter which can be considered either as a fixed constant x_a or as a lognormal probability distribution with mean $E[x_a]$ and standard deviation $STD[x_a]$.

In the remainder of this report, we will refer to the model solution simply as $n(t, x; \theta)$ where $\theta \subset \mathbb{R}^p$ is a set of parameters which describes the model. (This includes the case that x_a is described by a probability measure, where $\eta(t, x)$ was used in the previous exposition.) This is done to simplify notation, and it will always be clear from context which parametrization is being used. Obviously, the value of p will vary depending upon the parametrization. The various possibilities are summarized in Table 3.

We now return to the OLS formulation (19) of the inverse problem,

$$\theta_{OLS} = \arg\min_{\theta \in \Theta} \sum_{k,j} \mathcal{R}_{kj}^2 = \arg\min_{\theta \in \Theta} \sum_{k,j} \left(I[\tilde{n}](t_j, z_k^j; \theta) - N_k^j \right)^2,$$

where now Θ is a closed bounded subset of $\mathbb{R}^p$. Using the data $\{n_k^j\}$ as realizations of the random variables $\{N_k^j\}$,

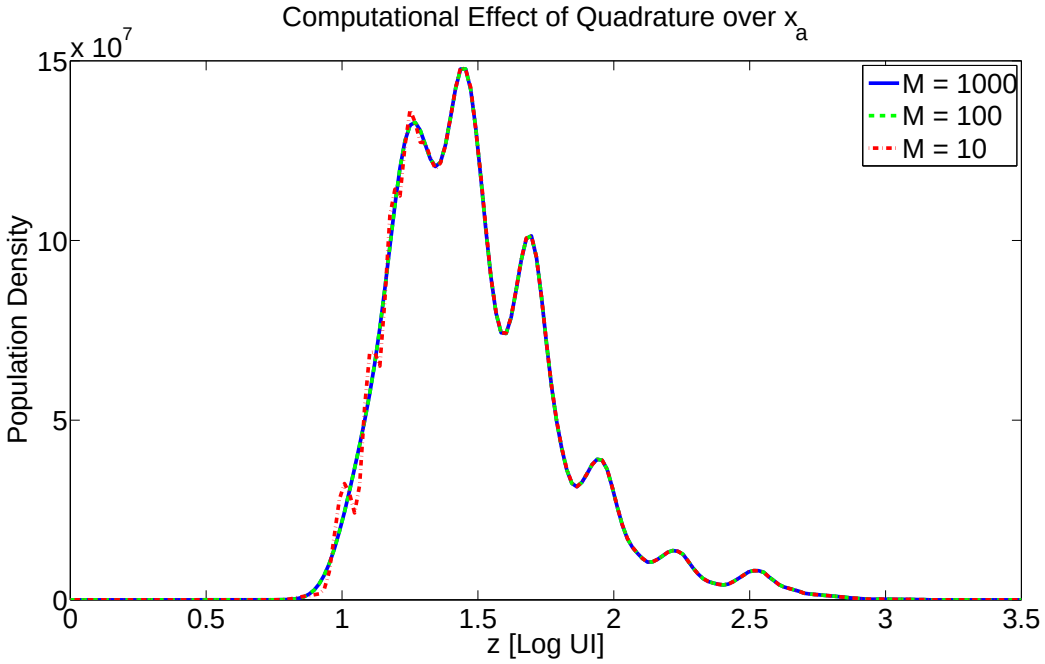

Figure 8: Effect of the number of nodes M used to approximate the total population density $\eta(t,x)$ in Equation (22). Using $M = 100$ seems more than sufficient.

we would like to compute the estimate

$$\hat{\theta}_{OLS}(\{n_k^j\}) = \arg\min_{\theta\in\Theta} J(\theta|\{n_k^j\}) = \arg\min_{\theta\in\Theta} \sum_{k,j} \left(I[\tilde{n}](t_j, z_k^j; \theta) - n_k^j\right)^2. \tag{25}$$

However, we have only an approximate numerical solution with which to compare the data. Thus we actually compute the approximate estimate

$$\hat{\theta}_{OLS}(h_t, N_x, M; \{n_k^j\}) = \arg\min_{\theta\in\Theta} J_A(\theta|\{n_k^j\}) = \arg\min_{\theta\in\Theta} \sum_{k,j} \left(I_A[\tilde{n}](t_j, z_k^j; \theta) - n_k^j\right)^2, \tag{26}$$

where we have now explicitly emphasized the dependence of the parameter estimate on the computational accuracy of the numerical solution. The continuous dependence of the model solution $\tilde{n}(t,z;\theta)$ on the parameter θ (regardless of which particular parametrization is used) follows easily from the method of characteristics solution of Section 2.1. Numerical convergence with respect to h_t, N_x, and M follow directly from well-known results regarding the trapezoidal rule for quadrature, linear interpolation of a smooth function, and the midpoint rule for quadrature, respectively. As such, it can be shown (see, e.g., the arguments of [15, Ch. 3] that the approximate estimates $\hat{\theta}_{OLS}(h_t, N_x, M; \{n_k^j\})$ will converge to some $\hat{\theta}^*_{OLS}$ which minimizes (25) as $N_x, M \to \infty$, and $h_t \to 0$. It should be noted that the possible nonuniqueness of the minimizer $\hat{\theta}^*_{OLS}$ is a common issue in inverse problems. We forgo techniques such as Tikhonov regularization in this report, choosing to focus instead on the accuracy of the best fit models $n(t,z;\hat{\theta}_{OLS})$ in fitting a particular data set, regardless of uniqueness (although these issues must be dealt with in order to establish standard errors, confidence intervals, etc.). For the remainder of this report, we will not distinguish between $\hat{\theta}_{OLS}$, $\hat{\theta}^*_{OLS}$, or $\hat{\theta}_{OLS}(h_t, N_x, M; \{n_k^j\})$. It should also be noted that this best-fit parameter, which is itself an estimate of the random variable θ_{OLS}, will be data-realization dependent. However, for a good model and a sufficiently large data set, $\hat{\theta}_{OLS}$ is an unbiased estimator of θ_{OLS} [20, 28, 62].

The optimization (26) has been implemented in MATLAB using the `fmincon` function, which is a variation of the BFGS-active set algorithm for bound-constrained parameters. The parameter constraints are summarized in Table 4.

Table 3: Summary of possible parameterizations for the compartmental model, with the set $\theta \in \mathbb{R}^p$ of parameters describing the model in each case.

Model	Parameters	p	Model	Parameters	p
A1B1	$\theta = \{x_a, c, k, \alpha_0, \alpha\}$	5	A1B1dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, \alpha\}$	6
A1B2	$\theta = \{x_a, c, k, \alpha_0, \alpha, \beta\}$	6	A1B2dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, \alpha, \beta\}$	7
A1B3	$\theta = \{x_a, c, k, \alpha_0, \alpha, \beta_0\}$	6	A1B3dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, \alpha, \beta_0\}$	7
A1B4	$\theta = \{x_a, c, k, \alpha_0, \alpha, \beta_0, \beta\}$	7	A1B4dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, \alpha, \beta_0, \beta\}$	8
A1B5	$\theta = \{x_a, c, k, \alpha_0, \alpha, \{\beta_i\}\}$	12	A1B5dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, \alpha, \{\beta_i\}\}$	13
A2B1	$\theta = \{x_a, c, k, \{\alpha_i\}\}$	9	A2B1dist	$\theta = \{E[x_a], STD[x_a], c, k, \{\alpha_i\}\}$	10
A2B2	$\theta = \{x_a, c, k, \{\alpha_i\}, \beta\}$	10	A2B2dist	$\theta = \{E[x_a], STD[x_a], c, k, \{\alpha_i\}, \beta\}$	11
A2B3	$\theta = \{x_a, c, k, \{\alpha_i\}, \beta_0\}$	10	A2B3dist	$\theta = \{E[x_a], STD[x_a], c, k, \{\alpha_i\}, \beta_0\}$	11
A2B4	$\theta = \{x_a, c, k, \{\alpha_i\}, \beta_0, \beta\}$	11	A2B4dist	$\theta = \{E[x_a], STD[x_a], c, k, \{\alpha_i\}, \beta_0, \beta\}$	12
A2B5	$\theta = \{x_a, c, k, \{\alpha_i\}, \{\beta_i\}\}$	16	A2B5dist	$\theta = \{E[x_a], STD[x_a], c, k, \{\alpha_i\}, \{\beta_i\}\}$	17
A3B1	$\theta = \{x_a, c, k, \alpha_0, t^*, \alpha\}$	6	A3B1dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, t^*, \alpha\}$	7
A3B2	$\theta = \{x_a, c, k, \alpha_0, t^*, \alpha, \beta\}$	7	A3B2dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, t^*, \alpha, \beta\}$	8
A3B3	$\theta = \{x_a, c, k, \alpha_0, t^*, \alpha, \beta_0\}$	7	A3B3dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, t^*, \alpha, \beta_0\}$	8
A3B4	$\theta = \{x_a, c, k, \alpha_0, t^*, \alpha, \beta_0, \beta\}$	8	A3B4dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, t^*, \alpha, \beta_0, \beta\}$	9
A3B5	$\theta = \{x_a, c, k, \alpha_0, t^*, \alpha, \{\beta_i\}\}$	13	A3B5dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, t^*, \alpha, \{\beta_i\}\}$	14
A4B1	$\theta = \{x_a, c, k, \alpha_0, t^*, \{\alpha\}\}$	10	A4B1dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, t^*, \{\alpha\}\}$	11
A4B2	$\theta = \{x_a, c, k, \alpha_0, t^*, \{\alpha\}, \beta\}$	11	A4B2dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, t^*, \{\alpha\}, \beta\}$	12
A4B3	$\theta = \{x_a, c, k, \alpha_0, t^*, \{\alpha\}, \beta_0\}$	11	A4B3dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, t^*, \{\alpha\}, \beta_0\}$	12
A4B4	$\theta = \{x_a, c, k, \alpha_0, t^*, \{\alpha\}, \beta_0, \beta\}$	12	A4B4dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, t^*, \{\alpha\}, \beta_0, \beta\}$	13
A4B5	$\theta = \{x_a, c, k, \alpha_0, t^*, \{\alpha\}, \{\beta_i\}\}$	17	A4B5dist	$\theta = \{E[x_a], STD[x_a], c, k, \alpha_0, t^*, \{\alpha\}, \{\beta_i\}\}$	18
A5B1	$\theta = \{x_a, c, k, \{a_i^{(p)}\}\}$	21	A5B1dist	$\theta = \{E[x_a], STD[x_a], c, k, \{a_i^{(p)}\}\}$	22
A5B2	$\theta = \{x_a, c, k, \{a_i^{(p)}\}, \beta\}$	22	A5B2dist	$\theta = \{E[x_a], STD[x_a], c, k, \{a_i^{(p)}\}, \beta\}$	23
A5B3	$\theta = \{x_a, c, k, \{a_i^{(p)}\}, \beta_0\}$	22	A5B3dist	$\theta = \{E[x_a], STD[x_a], c, k, \{a_i^{(p)}\}, \beta_0\}$	23
A5B4	$\theta = \{x_a, c, k, \{a_i^{(p)}\}, \beta_0, \beta\}$	23	A5B4dist	$\theta = \{E[x_a], STD[x_a], c, k, \{a_i^{(p)}\}, \beta_0, \beta\}$	24
A5B5	$\theta = \{x_a, c, k, \{a_i^{(p)}\}, \{\beta_i\}\}$	28	A5B5dist	$\theta = \{E[x_a], STD[x_a], c, k, \{a_i^{(p)}\}, \{\beta_i\}\}$	29

Table 4: Summary of bound-constraints for the OLS parameter estimation problem (26). The parameters $\{\alpha_i\}$, $\{a_i^{(p)}\}$, and $\{\beta_i\}$ must be positive. The feasibility of the remaining bounds has been determined computationally.

Parameter	Minimum	Maximum
x_a	1	20
$E[x_a]$	5	12
$STD[x_a]$	0	4
c	1×10^{-4}	1×10^{-2}
k	0	1×10^{-3}
$\{\alpha_i\}$ or $\{a_i^{(p)}\}$	0	1
$\{\beta_i\}$	0	1

3.5 Information Theoretic Model Selection

Each possible parametrization presented thus far gives rise to a distinct mathematical model which can be fit to a data set in the prescribed manner. Based upon the results for each model, we would like to determine which parametrization is most appropriate and use those results to draw conclusions regarding division-linked and/or longitudinal changes in the behavior of the cell culture. In order to do this, we must establish some formal mechanism which permits the objective comparison of different models.

One common approach is hypothesis testing for model refinements [8, 11]. However, such methods are only useful for pairwise comparisons, and are better suited for comparison against an experimental control [23]. Moreover, such methods do not apply unless one of the two models in the comparison is contained within the other model (for instance, our parametrization **B4** contains **B3** as a special case). While several parameterizations discussed in this document are indeed contained within other parameterizations, this is not universally the case (e.g., there is no containment relationship between **B2** and **B3**).

A more general approach, based upon the premises of information theory, is found in the Akaike Information Criterion (AIC). Briefly, for models with independent, homoscedastic, normally distributed errors, it can be shown that (recall that p is the number of parameters estimated)

$$AIC = m \log\left(\frac{J(\hat{\theta}_{OLS})}{m}\right) + 2p, \tag{27}$$

where m is the total number of data points, is an approximately unbiased estimate of the "expected relative Kullback-Leibler distance" (information loss) when a model is used to describe a data set [23]. Given a set of R models, the AIC can be computed for each model; we seek the model which results in the smallest AIC value. It should be emphasized that the AIC is only an estimate of information loss, and this estimate depends on the particular data set being used. (When comparing different models with the AIC, the same data set must be used to fit each model.) As discussed in Section 3.1, we cannot ascertain a priori that the measurement errors are normally distributed with constant variance. However, the use of an OLS framework already constitutes an assumption of homoscedasticity. The assumption of normality does not seem to be a significantly greater burden, and we proceed with the AIC in spite of these issues, recognizing that the use of the AIC will be only suggestive. As will be shown in Section 4, there is a very clear preference among the models when ranked by the AIC. As such, we do not expect our results to change significantly for a different error model. Similarly, the derivation of the AIC assumes that any model which is fit to the data is sufficiently accurate so that the assumption $E[N_k^j] = n(t_j, x_j; \hat{\theta}_{OLS})$ is valid. While this assumption may break down for the least accurate of the models tested in this report (see Section 4), it is a standard assumption in the OLS framework (19), provided the estimate $\hat{\theta}_{OLS}$ is sufficiently close to θ_0 for each particular model.

There is an element of parsimony in the AIC, as a model which fits the data poorly (high $J(\hat{\theta}_{OLS})$) or which contains a large number of parameters (high p) will have a comparatively larger AIC. Yet, rather than using the AIC to determine a single 'best' model, additional theory is available. If AIC_{min} is the smallest computed AIC value, then we can define the AIC differences

$$\Delta_r = AIC_r - AIC_{min}, \tag{28}$$

for $1 \leq r \leq R$, where AIC_r is the AIC value computed when model r is fit to the data. Finally, we can compute the Akaike weights

$$w_r = \frac{\exp\left(\frac{-\Delta_r}{2}\right)}{\sum_r \exp\left(\frac{-\Delta_r}{2}\right)}. \tag{29}$$

It can be shown (either by likelihood ratio tests or in a Bayesian framework, see [23]) that the AIC weight w_r can be interpreted as the probability that model r is the best model to describe the data (given the set of R possible models). Thus, after each model from Table 3 is fit to a data set, we can compute the Akaike weights for the set of candidate models and use these to assess the necessity of various mathematical features (e.g., division dependence of cell death rates) in describing the data. A complete derivation of the AIC and Akaike weights, as well as numerous examples and exhaustive references, can be found in [23].

4 Results and Discussion

The model calibration results for each possible parametrization of the compartmental model considered in this report are summarized in Table 5. The approximate OLS costs $J_A(\hat{\theta}_{OLS})$ are shown for each parametrization, as well as the computed AIC values and AIC differences. The models are also ranked in terms of their relative information theoretic loss.

The AIC selected model is parametrization A5B5 with lognormally distributed AutoFI (henceforth, A5B5dist) with a cost $J_A(\hat{\theta}_{OLS}) = 3.0535 \times 10^{11}$. This parametrization resulted in a model with not only the smallest AIC value, but also the lowest cost (meaning that the decrease in cost more than offset the additional parameters). The optimal solution for parametrization A5B5dist is depicted in comparison to the data in Figure 9. The estimated piecewise linear proliferation rates can be found in Figure 10, and the estimated death rates are summarized

Table 5: Summary of results for the various models considered in this report. The AIC-selected best model, parametrization A5B5 with lognormally distributed AutoFI, not only has the lowest cost, the OLS cost of this model is so much smaller (compared to the other models tested) that its AIC value is *significantly* lower than for any other model. The Akaike weights are not shown, as the weight assigned to model A5B5dist must be greater than $1 - 50\exp(-43/2) > 1 - 1 \times 10^{-7}$.

Model	$J_A(\hat{\theta}_{OLS})$	AIC_r	Δ_r	Rank	Model	$J_A(\hat{\theta}_{OLS})$	AIC_r	Δ_r	Rank
A1B1	48.9309×10^{11}	89459	11850	50	A1B1dist	44.5493×10^{11}	89059	11450	45
A1B2	48.5765×10^{11}	89430	11821	49	A1B2dist	26.0134×10^{11}	86753	9144	31
A1B3	46.0968×10^{11}	89205	11596	48	A1B3dist	18.9754×10^{11}	85400	7791	26
A1B4	46.0439×10^{11}	89202	11593	47	A1B4dist	18.9754×10^{11}	85402	7793	27
A1B5	35.5868×10^{11}	88107	10498	40	A1B5dist	17.9868×10^{11}	85183	7574	24
A2B1	30.8384×10^{11}	87487	9878	37	A2B1dist	42.3075×10^{11}	88845	11236	43
A2B2	30.4566×10^{11}	87436	9827	36	A2B2dist	21.2095×10^{11}	85886	8277	29
A2B3	28.6677×10^{11}	87176	9567	33	A2B3dist	14.8563×10^{11}	84359	6750	16
A2B4	28.6677×10^{11}	87178	9569	34	A2B4dist	14.8562×10^{11}	84361	6752	17
A2B5	28.6677×10^{11}	87188	9579	35	A2B5dist	14.8562×10^{11}	84371	6762	18
A3B1	45.4019×10^{11}	89140	11531	46	A3B1dist	42.9086×10^{11}	88900	11291	44
A3B2	37.3875×10^{11}	88309	10700	42	A3B2dist	11.9759×10^{11}	83428	5819	11
A3B3	34.8434×10^{11}	88007	10398	38	A3B3dist	13.5090×10^{11}	83945	6336	12
A3B4	34.8376×10^{11}	88008	10399	39	A3B4dist	10.5215×10^{11}	82875	5266	10
A3B5	18.7334×10^{11}	85357	7748	25	A3B5dist	6.9142×10^{11}	81084	3475	8
A4B1	25.3453×10^{11}	86648	9039	30	A4B1dist	36.6830×10^{11}	88236	10627	41
A4B2	16.8159×10^{11}	84890	7281	22	A4B2dist	5.5690×10^{11}	80152	2543	7
A4B3	17.1422×10^{11}	84973	7364	23	A4B3dist	8.3562×10^{11}	81893	4284	9
A4B4	16.6846×10^{11}	84859	7250	21	A4B4dist	5.0699×10^{11}	79752	2143	6
A4B5	16.4652×10^{11}	84812	7203	20	A4B5dist	4.5712×10^{11}	79318	1709	5
A5B1	19.6228×10^{11}	85572	7963	28	A5B1dist	27.3143×10^{11}	86993	9384	32
A5B2	15.0638×10^{11}	84440	6831	19	A5B2dist	3.5086×10^{11}	78193	584	4
A5B3	14.6710×10^{11}	84327	6718	13	A5B3dist	3.2607×10^{11}	77879	270	3
A5B4	14.6740×10^{11}	84330	6721	14	A5B4dist	3.0918×10^{11}	77653	43	2
A5B5	14.6727×10^{11}	84339	6730	15	A5B5dist	3.0535×10^{11}	77609	0	1

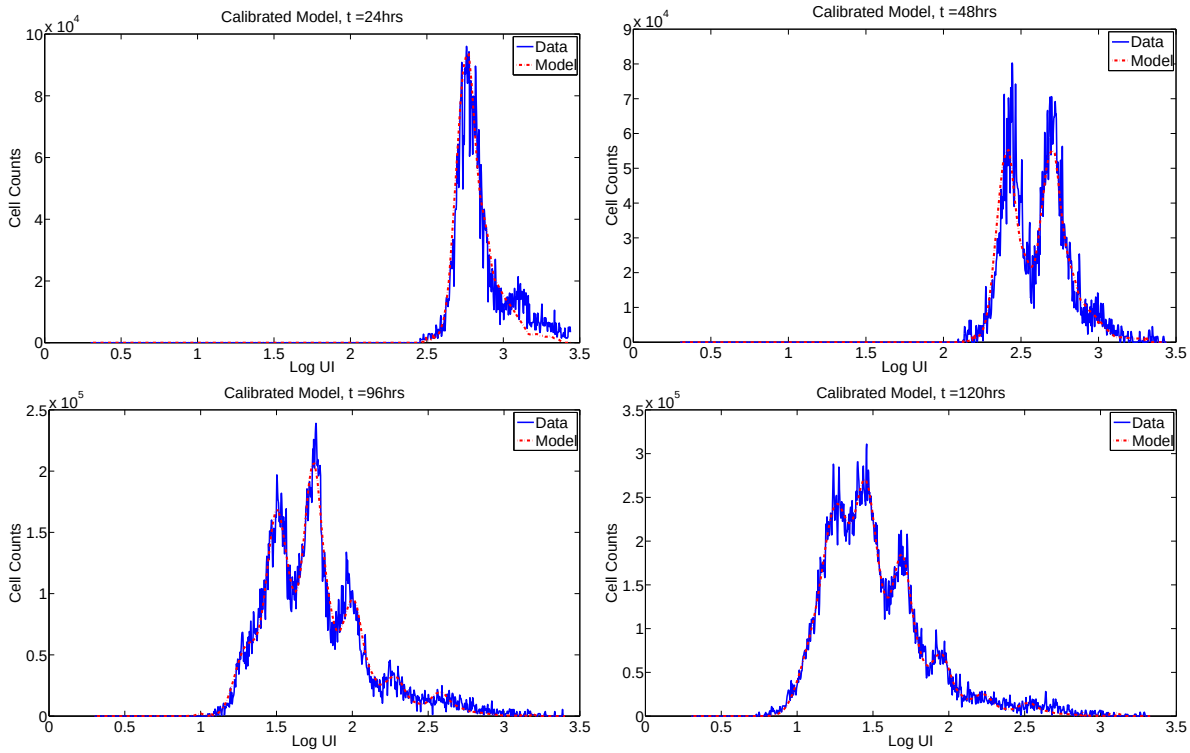

Figure 9: Best-fit solution $I_A[\tilde{n}](t, z; \hat{\theta}_{OLS})$ for parametrization A5B5dist. Total cost $J_A(\hat{\theta}_{OLS}) = 3.0535 \times 10^{11}$.

Table 6: Estimated death rates β_i in terms of the number i of divisions undergone.

Divisions	Death Rate (1/hr)
0	0.0165
1	0.0000
2	0.0000
3	0.0000
4	0.0012
5	0.0544
6	0.1572

in Table 6. For the AutoFI distribution, the best-fit lognormal distribution has mean $E[x_a] = 8.739$ UI and $STD[x_a] = 3.534$ UI. The estimated Gompertz label decay parameters are $c = 5.641 \times 10^{-3}$ and $k = 1 \times 10^{-9}$.

As can clearly be seen in Figure 9, the compartmental model (with suitable parametrization) is capable of accurately describing the particular data set used for model calibration in this report. The most notable shortcoming of the model occurs at $t = 24$ hours, where a distinct cohort of cells with high CFSE FI can be seen in the data and is not modeled accurately. As discussed in [18], this cohort is believed to be either cell duplets or some other anomalous cell types which were not properly gated out of the measured cell data, and such cells should not be an issue in future data sets. It also appears that neither of the two generations in the model solution at $t = 48$ hours contains enough cells (when compared to the data at that time). This may also be partly explained as a systematic error resulting from the presence of cell duplets in the data. It is also possible that small errors associated with the manner in which counted beads (see [66, Ch. 1]) are used to determine the total population size.

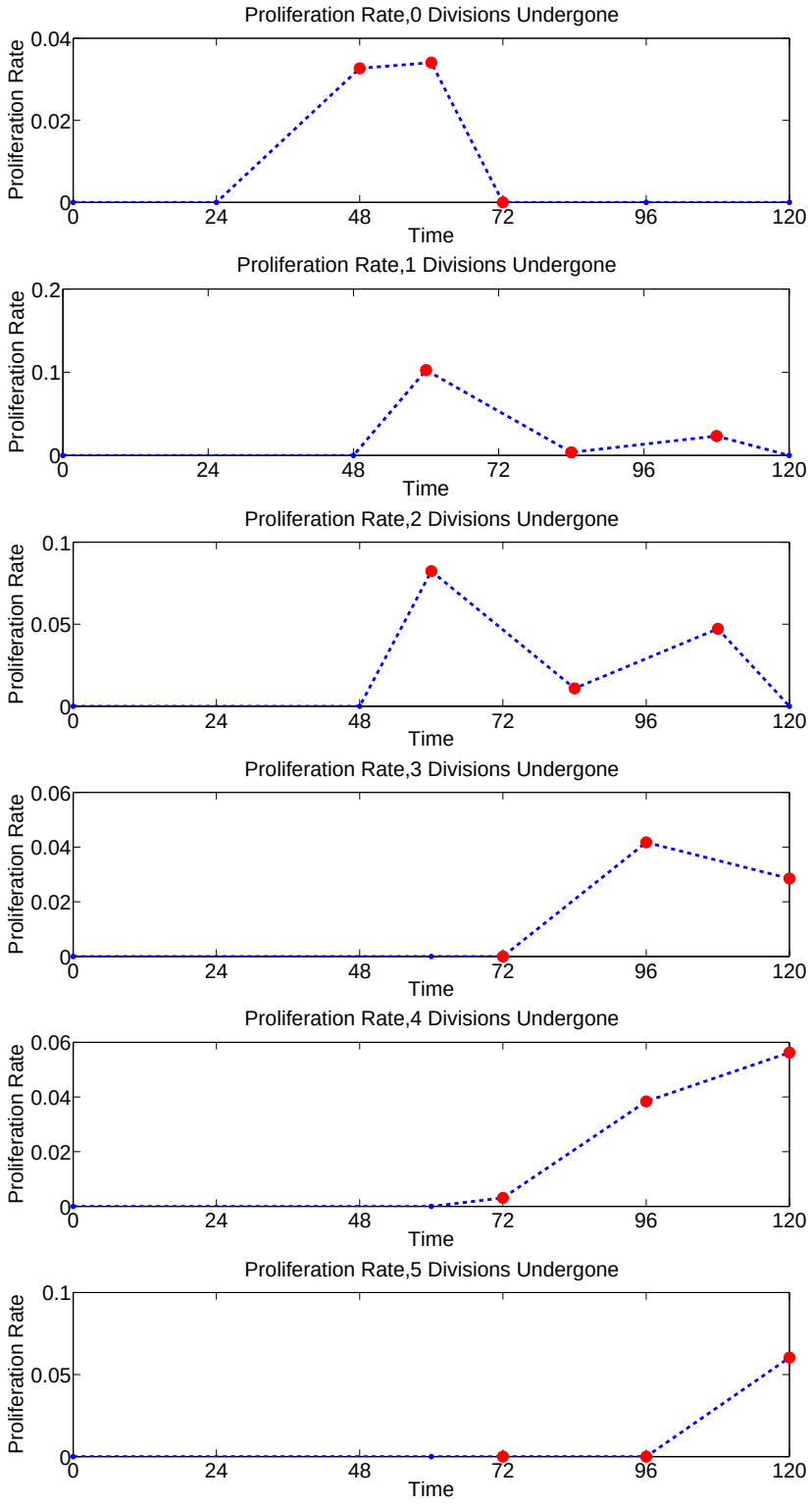

Figure 10: OLS best-fit piecewise linear proliferation rate functions for each division number. Red circles indicate nodes which were estimated in the inverse problem.

One of the primary goals of considering various parameterizations for the proliferation and death rates (Section 3) was to investigate the dependence of these rates on division number and on time. The best-fit parametrization A5B5dist features a proliferation rate which depends both on time and division number, as well as a death rate which depends on division number. Additionally, the AutoFI parameter x_a is lognormally distributed, which was not considered in previous efforts [18]. Given the overwhelming weight assigned to the parametrization A5B5dist in the information theoretic framework, it is tempting to conclude that each of these features is necessary in accurately modeling the data. Because the data set contains 4289 points, even a small difference in OLS cost (compare, for example, parameterizations A5B5dist and A5B4dist) results in significantly different AIC values. However, the AIC (27) is derived under the assumptions of independent, homoscedastic, normally distributed errors. If these assumptions are not valid, particularly if the 4289 points are not independent, then the magnitude of the AIC differences may be misleadingly large.

In spite of these potential setbacks, there are still several useful conclusions which can be safely drawn. As expected, the worst parameterizations (in terms of both OLS cost and AIC rank) are those which do not permit cell death in the population (**B1**). Parameterizations which feature probabilistically distributed AutoFI are more accurate than parameterizations which use a constant parameter x_a to describe AutoFI. Among the models which use a constant parameter x_a to describe AutoFI, the most parsimonious model (that is, the AIC selected model) is parametrization A5B4. (Parameterizations A5B4 and A5B5 differ minimally in cost, but A5B4 has fewer parameters.) The best-fit solution for this model is shown in comparison to the data in Figure 11. *We find that a model which fails to account for variability in AutoFI in the population of cells does not adequately describe the increasing heterogeneity of the population of cells as division number increases.* This is particularly noteworthy for cells having undergone 4 or more divisions, where AutoFI constitutes a comparatively larger fraction of the measured FI of the cells. Such an observation has important experimental ramifications for the design of intracellular dyes. While it has long been known that a population of cells must obtain a high level of FI (relative to their AutoFI) during the initial staining process in order for the experimenter to resolve multiple rounds of division in the population [54, 60], we now see that the variability of AutoFI in the population of cells also has an effect on the peak-to-peak resolution of the data. While AutoFI is a property of the cells being measured (it arises from intracellular molecules which emit light in the frequency bands used to detect the intracellular dye), *focus may possibly be directed toward the design of dyes with spectral properties that minimally overlap with common intracellular molecules.*

As in previous work [18, 19], we find that time dependence is a significant feature of the proliferation rates, given the model formulation (6). Significantly, we find that the population of cells cannot be accurately modeled by considering only a delay in the time to first division. For instance, the calibrated model using parametrization A4B5dist (which is the AIC selected model among those which does not feature completely time dependent proliferation) is shown in Figure 12. This parametrization does not permit any proliferation until $t \geq t^*$, thus enforcing a delay before any division occurs in the population. Even with this feature, subsequent divisions of cells emerge too quickly in the model solution. Thus more complex time-dependence (parametrization **A5**) appears to be necessary, as the resulting decrease in cost outweighs the additional parameters.

The compartmental model was motivated by a desire to compute quantities such as cell numbers from the best-fit model solution. As noted above, previous methods for obtaining cell numbers relied on some form of deconvolution of the histogram data, typically via fitting by a series of normal or lognormal curves. While the compartmental model is more mathematically involved and requires considerably more time for fitting to data (a few minutes to a few hours, depending upon the parametrization used and the accuracy of the initial parameter guess for the BFGS algorithm), it does not require any assumption as to the shape of the distribution of cells within a single generation. Given a calibrated model solution $n(t, x; \hat{\theta}_{OLS})$, one can compute the total number of cells

$$N_i(t) = \int_{x_a}^{\infty} n_i(t, x; \hat{\theta}_{OLS}) dx \tag{30}$$

for each generation. It may also be of experimental interest to consider the number of precursors in the population. Because each cell division results in the formation of two daughter cells from a single mother cell, one must renormalize (by a factor of 2) the total number of cells in each generation in order to accurately analyze the proportion of cells proceeding through a specified number of divisions. Precursors, then, are cells in the original population (that is, at $t = 0$ hours) which eventually give rise to other cells with higher division numbers at later

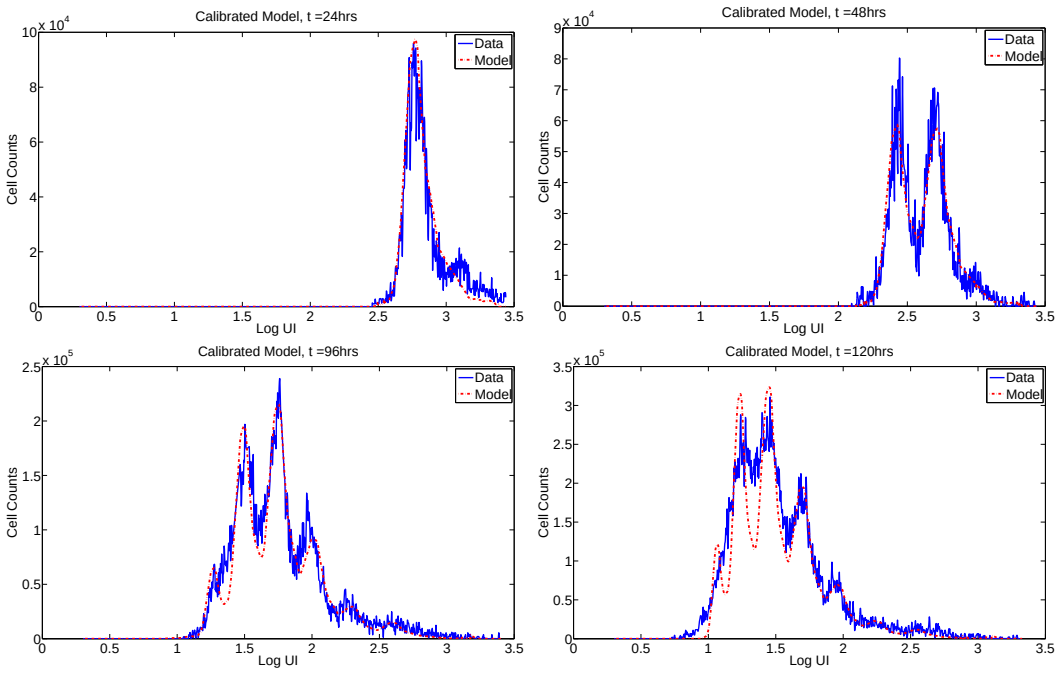

Figure 11: Among the models which do not use a lognormal distribution to describe AutoFI, the AIC selected model is parametrization A5B4. When comparing the best-fit solution to the data, it is clear that a model lacking an AutoFI distribution will result in peaks which are too distinct when compared to the data.

times. The number of precursors is

$$P_i(t) = \frac{N_i(t)}{2^i} = \frac{1}{2^i}\int_{x_a}^{\infty} n_i(t, x; \hat{\theta}_{OLS})dx. \tag{31}$$

Given that precursors represent numbers of cells in the original population, it follows that the total number of precursors

$$P(t) = \sum_{i=0}^{i_{\max}} P_i(t)$$

cannot increase in time (but may decrease as a result of cell death). Cell numbers and precursor numbers have been computed from the best-fit model solution (parametrization A5B5dist) and are shown in Figure 13 for undivided cells ($N_0(t)$, $P_0(t)$) and divided cells ($\sum_{i=1}^{i_{\max}} N_i(t)$, $\sum_{i=1}^{i_{\max}} P_i(t)$) as well as total cells. It follows that such curves could easily be used to determine such parameters as approximate doubling time for the population, or the fraction of cells which do not divide. These parameters may be of particular importance in accounting for changes in behavior as a function of experimental conditions (e.g., strength of stimulation) or in a diagnostic setting.

5 Discussion

In this report, a label structured system of PDEs for a population of dividing cells, indexed by the number of divisions undergone, is derived and fit to data. Under the appropriate assumptions for the label loss rate, autofluorescence parameter, proliferation rates, and death rates, such a model can accurately fit an experimental data set (Figure 9). Because each generation of cells is mathematically described by a separate structured density function, the proliferation and death rates can be estimated directly in terms of division number, and there is

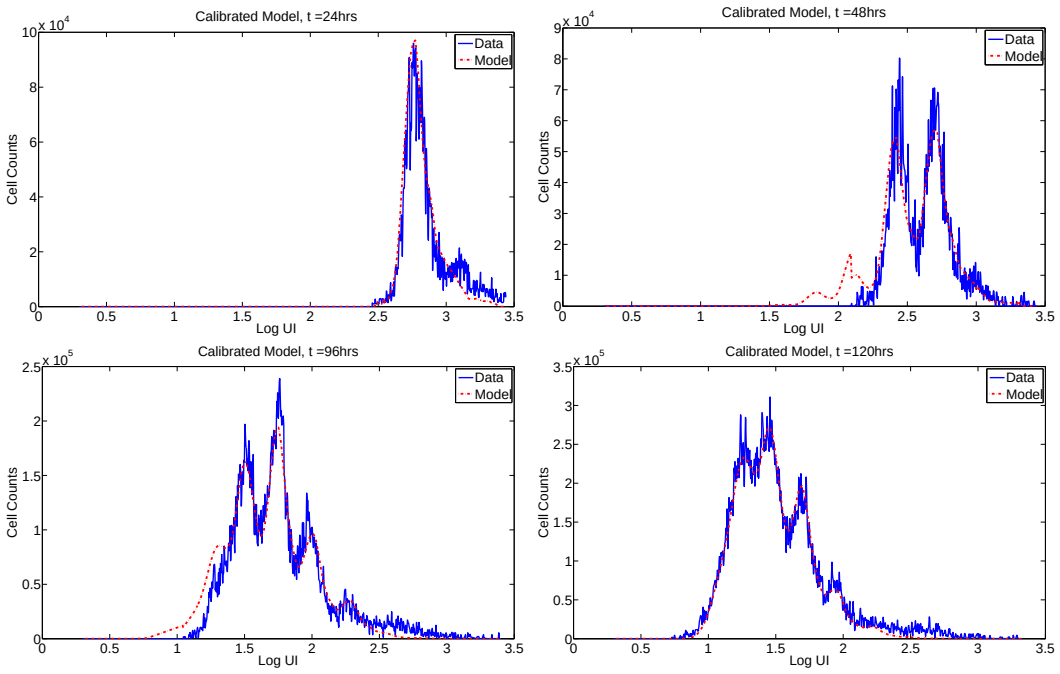

Figure 12: Best-fit model solution with parametrization A4B5dist, the AIC selected model among the subset of models which does not feature completely time-dependent proliferation. While this parametrization includes a delay before the first division is reached, this is still insufficient to describe the data as cells proceed through subsequent rounds of division too quickly. The discontinuity in the model solution at $t = 48$ hours is a result of a sudden change in the size of the histogram bins on which the data is specified.

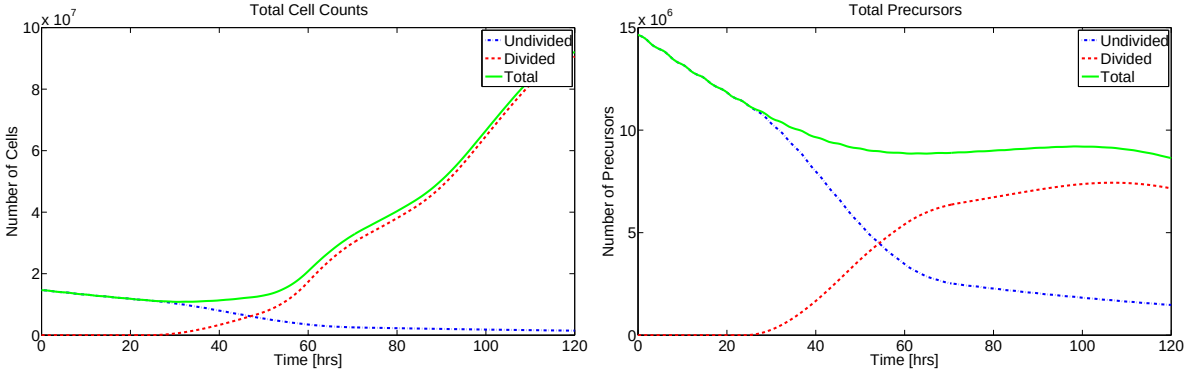

Figure 13: Total cell counts (left) and total precursors (right) in terms of undivided cells, divided cells, and total cells. The values are computed from the best-fit model solution $n(t, x; \hat{\theta}_{OLS})$ with parametrization A5B5dist. Numerical values at data collection times are summarized in Tables 7 and 8. The slight increase (less than 0.3%) in the total number of precursors between $t \approx 60$ hours and $t \approx 90$ hours is within the range of numerical error for the computed solution.

no need for any parametrization of these rates in the structure variable. This is in contrast to the previous fragmentation model (1) from [18]. The AIC-selected best-fit compartmental model contains 29 parameters and results in a best-fit OLS cost of $J_A(\hat{\theta}_{OLS}) = 3.0535 \times 10^{11}$ while the best-fit fragmentation model contained

73 parameters and resulted in a cost of 3.0901×10^{11}. Thus the compartmental model appears quite superior to the previous fragmentation model, as it contains fewer parameters and has a lower OLS cost. Additionally, the compartmental model can be used to compute cell numbers in terms of the number of divisions undergone. Certainly it may be possible to decrease the number of parameters in the fragmentation model by changing the placement of the nodes used for the estimation of the proliferation and death rate functions. Yet even if the total number of parameters could be decreased significantly without increasing the OLS cost of the fragmentation model, it still could not be used directly to compute cell numbers.

It is interesting to note that using the compartmental model we have found variability in AutoFI to be an essential feature of an accurate mathematical model. Yet the fragmentation model assumes only a constant value of AutoFI without significant sacrifice in accurately fitting the data (see [18]). The explanation for this unusual observation is the manner in which the proliferation rate is parameterized as a function of the structure variable (or the 'translated variable') in the fragmentation model. The large number of nodes used for the structural dependence of that proliferation rate (13 nodes) allows for significant variability in the proliferation rate, even among cells which are sufficiently close in the structural coordinate. Because the Gompertz function for label decay assumes that the rate of FI loss is directly proportional to the quantity of FI, a group of cells which divides immediately and then pauses will lose less label than a group of cells which waits for some time and then divides. In other words, the variability of the proliferation rate induces a variability in the label loss rate. As a consequence of this observation, it would be interesting to compare the effects of probabilistically distributed AutoFI with the effects of probabilistically distributed label loss rates in the compartmental model.

The major advantage of the compartmental model over previous efforts is the ability to compute cell numbers directly from the model solution (Figure 13). Because the compartmental model can be used to estimate the numbers of cells (or precursors) having undergone a specified number of divisions, biologically meaningful parameters can be assessed directly in terms of division number. For instance, the total number of precursors in the population, as a fraction of the original number, provides a meaningful estimation of cell viability. The total number of cells in the population can be used to estimate the population doubling time. As more complex experiments are conducted, the compartmental model could be easily generalized to account for division-linked changes (surface marker expression/differentiation, genetic mutations, etc.). Such features should be useful when comparing results from different data sets, such as when attempting to quantify the effects of a given chemical reagent, or distinguishing between diseased and healthy cells.

Of course, the meaningful comparison of parameter estimates between multiple data sets and experimental conditions relies upon quantification of the levels of uncertainty in the estimated parameters. This quantification, typically in the form of confidence bounds, is premised upon the accurate specification of the statistical model (17) which links the model to the data. In this report, the model was fit to the data in an ordinary least squares sense, with the tacit assumption that the error random variables $\mathcal{E}_{kj}$ have mean zero and constant variance. However, this assumption is not an accurate description of the data. While the misspecification of the statistical error model does not invalidate the ability of the compartmental model to (qualitatively) fit the available CFSE data set, it does impede the meaningful quantification of uncertainty in the parameter estimates. Work is ongoing to establish a suitable mathematical form for the statistical model accurately linking the histogram data to the model.

5.1 Observation and Error Models

It is hoped that the compartmental model will provide a quantitative framework for the comparison of data sets arising from cells in various biological and experimental conditions. However, before such a framework can be established, there is a need for meaningful confidence intervals to quantify the certainty with which individual parameters are estimated. This, in turn, relies upon an accurate statistical model for the CFSE histogram data.

Recall from Section 3.1 the assumption that the data is accurately described by the statistical model

$$N_k^j = I[\tilde{n}](t_j, z_k^j; \theta_0) + \mathcal{E}_{kj}, \tag{32}$$

where $\mathcal{E}_{kj}$ are independent random variables satisfying $E[\mathcal{E}_{kj}] = 0$. Then the best-fit parameter estimate is

$$\hat{\theta}_{WLS}(\{n_k^j\}) = \arg\min_{\theta\in\Theta} \sum_{k,j} \frac{r_{kj}^2}{w_{kj}} = \arg\min_{\theta\in\Theta} \sum_{k,j} \frac{1}{w_{kj}} \big(I[\tilde{n}](t_j, z_k^j; \theta) - n_k^j\big)^2, \tag{33}$$

where the residuals r_{kj} are realizations of the error random variables $\mathcal{E}_{kj}$. In theory, the weights w_{kj} are chosen to account for the variance of the $\mathcal{E}_{kj}$ following the assumptions of the statistical model. Thus, the statistical model has direct implications for the estimated best-fit parameter, given the data.

Two possible variance models were considered in Section 3.1. First, a constant variance (CV) statistical model was considered, in which case $Var(\mathcal{E}_{kj}) = \sigma_0^2$ and $w_{kj} = 1$ for all k and j. This results in the Ordinary Least Squares (OLS) framework (19). Second, a constant coefficient of variance (CCV) statistical model was considered, in which $Var(\mathcal{E}_{kj}) = \sigma_0^2(I[\tilde{n}](t_j, z_k^j; \theta_0))^2$ and $w_{kj} = (I[\tilde{n}](t_j, z_k^j; \theta_0))^{-2}$. This results in the Generalized Least Squares (GLS) framework (20). In the absence of any a priori knowledge regarding the correct form of the statistical model, the computationally simpler OLS model was used for the inverse problem.

In addition to the implications for confidence interval calculation discussed above, an accurate statistical model has implications for the weights w_{kj} in the inverse problem formulation (32). Additionally, the computation of the AIC values (Section 3.5) for model ranking and selection is premised upon modeling errors which are independent and normally distributed with constant variance. There is significant value, then, in ascertaining the properties of the error random variables and assessing the reliability of the assumptions made in the inverse problem procedure.

The residuals r_{kj} which result from fitting the model to the data are realizations of the random variables $\mathcal{E}_{kj}$. As explained in [20], the reliability of the statistical error model can be assessed by plotting the residuals r_{kj} and the modified residuals $r_{kj}/I[\tilde{n}](t_j, z_k^j; \theta_0)$ in terms of the model values $I[\tilde{n}](t_j, z_k^j; \theta_0)$. (In practice, of course, one must insert an estimate $\hat{\theta}_{OLS}$ or $\hat{\theta}_{GLS}$ in the place of the unknown θ_0 and use the approximate integral operator $I_A[\tilde{n}]$.) If a CV error model is sufficient to explain the noise in the data, then the residuals r_{kj} will be randomly distributed when plotted against the model values, while the variance of the modified residuals will decrease as the magnitude of the model increases. Alternatively, if a CCV model is sufficient to explain the noise in the data, then the modified residuals $r_{kj}/I_A[\tilde{n}](t_j, z_k^j; \theta_0)$ will be randomly distributed, while the original residuals will grow with the magnitude of the model. These observations are summarized in a hypothetical example in Figure 14. More details regarding the choice of statistical error and its effects on the inverse problem can be found in [8, 20, 28, 62].

Figure 15 contains the residuals and the modified residuals plotted in terms of the computed model values for the AIC-selected best-fit model parametrization A5B5dist. (Technically speaking, one should plot the modified residuals $r_{kj}/I_A[\tilde{n}](t_j, z_k^j; \theta_{GLS})$. However, the computation of the parameter estimate θ_{GLS} is quite expensive, and the model values $I_A[\tilde{n}](t_j, z_k^j; \theta_{GLS})$ would change minimally.) When the residuals are plotted in terms of the value of the observed model solution there is a clear increase in the variance of the residuals as the size of the model increases, providing an indication that the assumption of CV errors may be incorrect. However, the residuals also lack the fan-like structure typical of CCV errors (Figure 14). When the modified residuals are plotted in terms of the magnitude of the observed model solution, the pattern is distinctly nonrandom. Thus, it appears that the true statistical model for the errors may lie somewhere between the CV model (OLS estimation) and the CCV model (GLS estimation), perhaps slightly closer to the CV model.

The assumption that the error random variables at each of the data points are independent may also be problematic. For instance, when the residual plots are separated in terms of measurement times (Figure 16) additional structure is noticeable in the residual plots when compared to Figure 15. The independence of the error random variables can be investigated with a scatterplot of the residuals r_{kj} (which are considered as realizations of those random variables) in terms of the previous residual $r_{(k-1)j}$. If the error random variables were truly independent, then such a scatterplot would have no discernable structure. However, we see in Figure 17 that this is not the case, as there is a clear positive correlation between the sets of residuals.

There are two possibilities which may explain the positive correlation between the two sets of residuals. First, it is possible that neighboring data points are not independent. Because the data used to calibrate the model is histogram data, it is possible that the number of cells counted into adjacent bins (and hence, the error terms) might be linked by the location of the boundary separating the adjacent bins. In general, this might be demonstrated by investigating how the noise in the data changes as the bins used to generate the histogram data changes. Unfortunately, the data set used in this report was received with the bins already fixed. Still, these effects deserve careful consideration and must be addressed in future work. Some research has also indicated that cells which descend from a common precursor may share certain traits and/or behaviors [41, 71]. It is unclear how such correlation might impact either the error terms or the mathematical model itself.

A second possibility is that the model, though close to the data (see, e.g., Figure 9) does not satisfy the

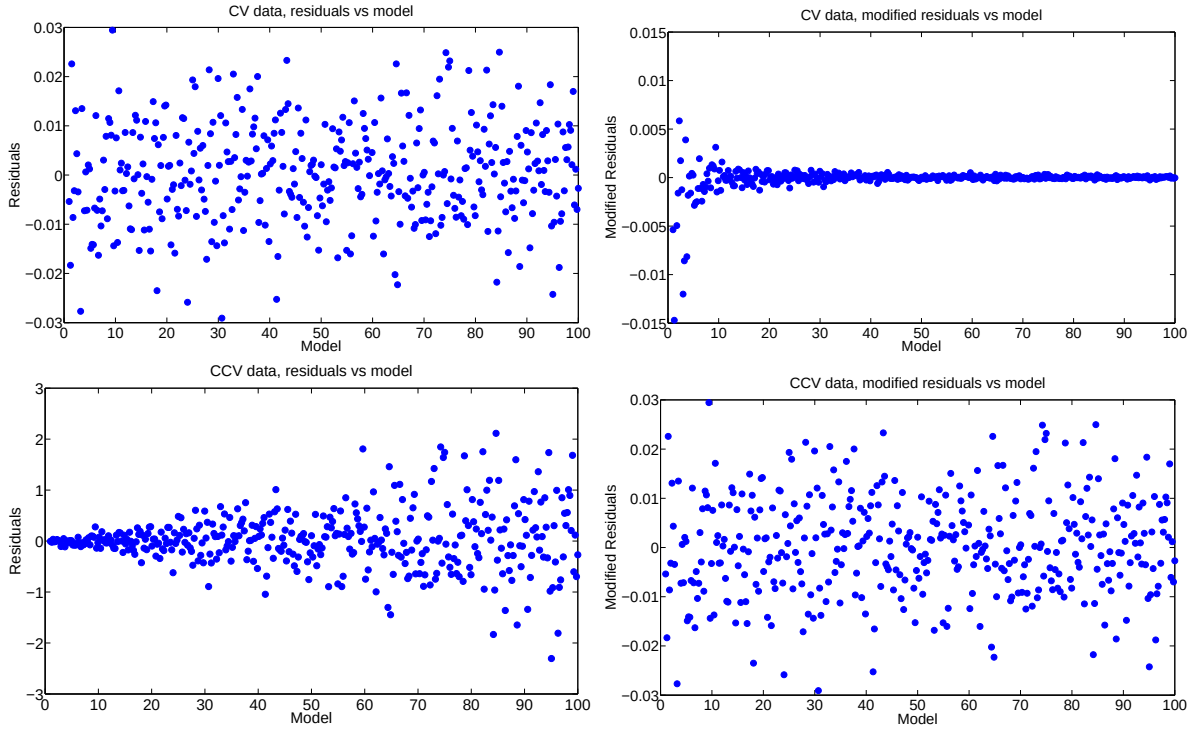

Figure 14: Top: Hypothetical residuals (left) and modified residuals (right) for constant variance (CV) data when plotted in terms of the model value. Bottom: Hypothetical residuals (left) and modified residuals (right) for constant coefficient of variation (CCV) data when plotted in terms of the model value. When the correct statistical model is used (top-left and bottom-right), the residuals (or modified residuals) appear randomly distributed. The fan-like structures in the top-right and bottom-left panels are characteristic of such residual plots when the statistical model for the measurement errors has been misspecified.

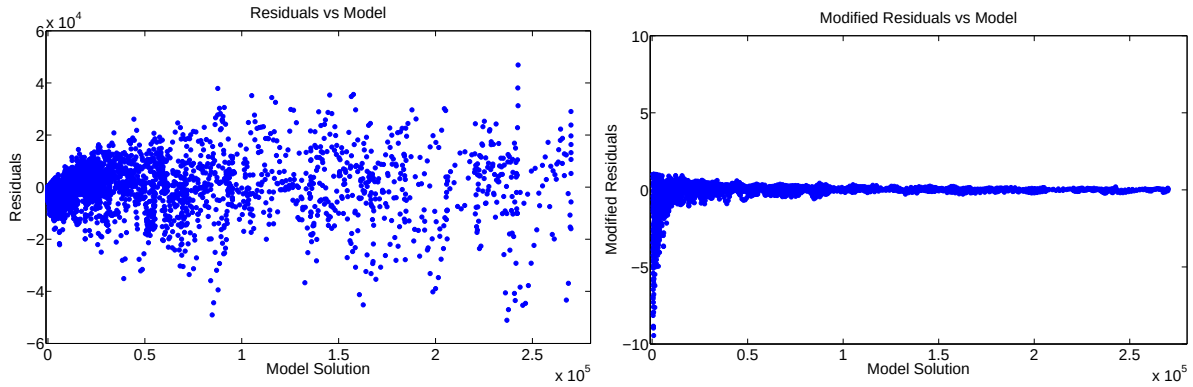

Figure 15: Residuals (left) and modified residuals (right) for the OLS best-fit model solution. Because neither graphic exhibits constant variance, the assumptions of CV error and CCV error must both be wrong. While the misspecification of the error term does not invalidate the ability of the compartmental model to fit the data, we cannot determine the statistical properties (e.g., confidence intervals) of the parameter estimates.

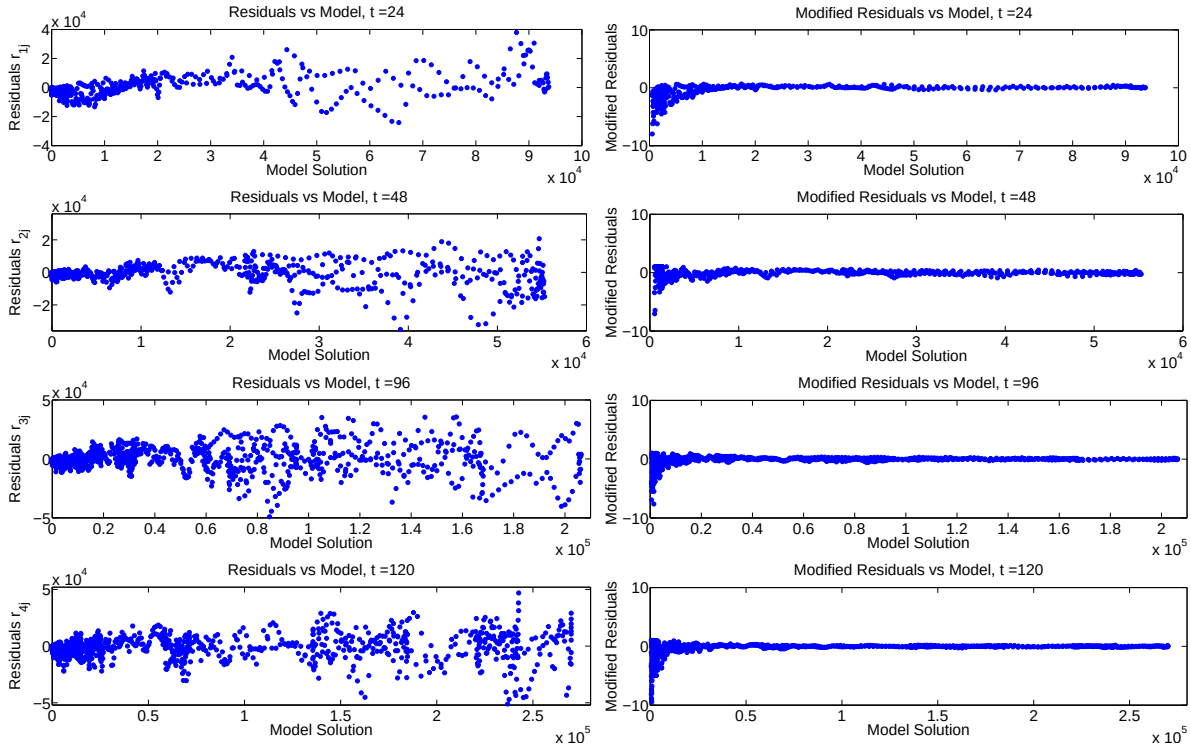

Figure 16: Residuals (left) and modified residuals (right) for the OLS best-fit model solution, shown separately for each measurement time. There is some additional structure which is evident in these residual plots which is not evident when the residuals for all measurement times are shown together (Figure 15). This has implications for the statistical model of the data.

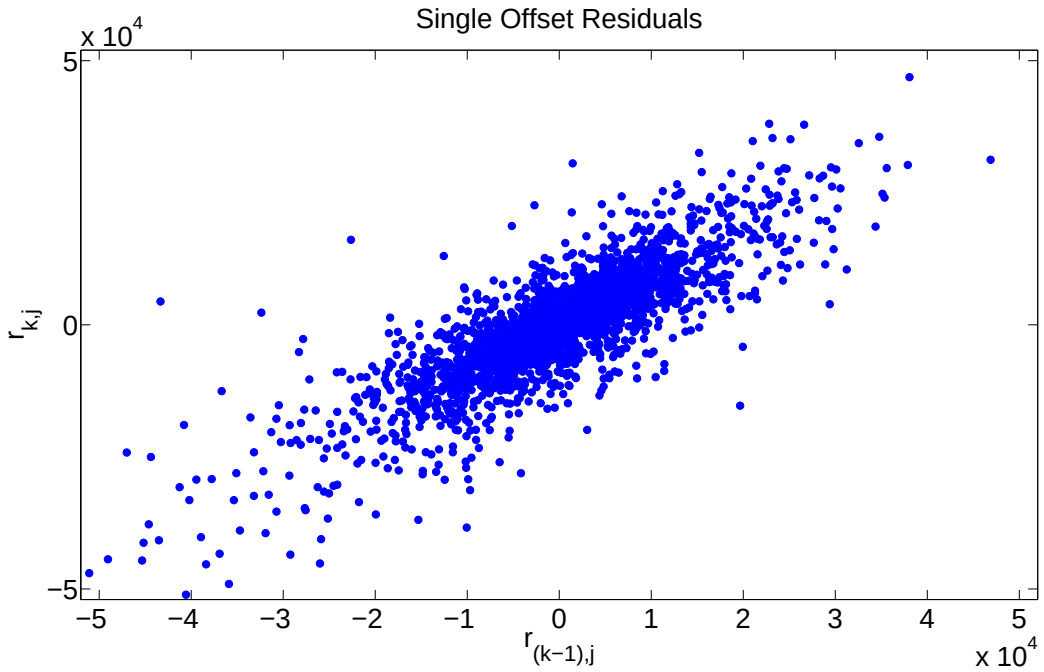

Figure 17: The assumption of independent errors for each observation can be checked by plotting the residuals r_{kj} against the offset residuals $r_{(k-1)j}$. If the errors were independent, there would be no discernable structure in such a graphic. However, we observe a clear positive correlation.

tacit assumption $E[N_k^j] = I[\tilde{n}](t_j, z_k^j; \theta_0)$ (or, equivalently, $E[\mathcal{E}_{kj}] = 0$). Given the convergence properties of the numerical solution discussed in Section 2.3, it seems unlikely that the failure of this assumption could be caused by any computational errors or approximations. A more likely explanation is the discrepancy between the assumptions regarding the collected data and the actual experimental reality. As discussed in [66, Ch. 1], the 5 samples of data collected (4 used for fitting the model plus one for the initial condition) are actually 5 separate samples taken from the same donor. While each sample receives an identical treatment, the assumption that all five samples are identical for all time may be inaccurate. Moreover, only a fraction of each sample is measured, and a scaling factor (which may also be subject to error) is used to adjust the resulting cell counts. Meanwhile, the model is derived under the assumption that each histogram represents a complete census of the cells in the population, and that the same population (i.e., cells arising from the same set of precursors) is measured each time. In order to correct for such a discrepancy between the assumptions of the mathematical model and the experimental reality, a more rigorous, detailed observation operator may be needed which accounts for the experimental sampling method, with its attendant sources of error. This may also help to resolve the slight negative bias of the residuals observed in Figure 15.

In fact, numerous additional data sets have been collected and have been studied qualitatively to determine which of these two possibilities is more likely. It is shown in [66, Ch. 4] that there is significant variability between otherwise identical samples (even when cells are taken from a single donor, treated identically, and measured at the same time). Thus the second possibility seems more likely, that the inaccuracy of the statistical model is the result of the failure of the statistical model to accurately incorporate sources of uncertainty which naturally arise in the experimental protocol (see [66, Ch. 1]). Ongoing work has been directed at understanding these additional sources of variability, along with the attendant implications for an accurate statistical model. Some preliminary results are given in [66, Ch. 4].

Table 7: Total numbers of cells in terms of division number. For each time and generation, the total number of cells has been computed from the OLS best-fit model solution (top), from a deconvolution of the data using normal curves (middle) and from a deconvolution of the data using lognormal curves (bottom). While the numbers computed from normal and lognormal curves are generally close together, there are clear differences between the values computed from the deconvolution methods and those obtained with the compartmental model. The most striking example occurs for $t = 120$ hours for cells having undergone 6 divisions. It is interesting to note that the division peaks in the histogram data are not well-resolved for such cells, making the accurate determination of cell numbers difficult.

	Divisions Undergone							
Time (hrs)	0	1	2	3	4	5	6	Total
	11339892	0	0	0	0	0	0	11339892
24	9211571	0	0	0	0	0	0	9211571
	9254681	0	0	0	0	0	0	9254681
	5881557	6555814	0	0	0	0	0	12437372
48	6298359	5473128	0	0	0	0	0	11771487
	6294945	5434570	0	0	0	0	0	11729515
	1906065	2478042	8092563	20976431	18520420	7588997	0	59562519
96	1970401	3284000	11019352	18184100	18307586	5252346	0	58017785
	2364520	3940800	10467773	17773515	18846649	5406993	0	58800249
	1476266	1978930	5605926	17086869	25529315	25986246	14337080	92000632
120	1969773	2969295	7881600	21017600	26272000	24958400	4597599	89666268
	2195762	3435673	7722653	17150087	26755781	29950079	5517118	92727154

5.2 Comparison with Deconvolution Techniques

Given the applicability of the compartmental model (once calibrated) to computing the numbers of cells having undergone a specified number of divisions, a relevant comparison can be drawn between the results of a label structured PDE model and the commonly used deconvolution techniques. In Table 7, the number of cells in each generation is computed at each measurement time. For each time and generation, the top number is computed from Equation (30). The middle number is computed by first fitting the function

$$\psi(z_k^j) = \sum_{i=0}^{i_{\max}} \psi_i(z_k; s_i, \mu_i, \sigma_i)$$

to the data at a given time, where $\psi_i(z_k; k_i, \mu_i, \sigma_i)$ is a normal density function with mean μ_i and standard deviation σ_i, scaled by a factor s_i. The method of fitting is ordinary least squares. Then the total number of cells having undergone i divisions is $\sum_k \psi_i(z_k; s_i, \mu_i, \sigma_i)$. The final number in each block of Table 7 is computed in an analogous manner, but with lognormal rather than normal density functions.

Unsurprisingly, the two deconvolution techniques (fitting with a series of normal or lognormal curves) provide estimates of cell numbers which are fairly consistent. However, these estimates occasionally differ from estimates obtained from the compartmental model. Of particular note is the difference for cells having undergone 6 divisions at $t = 120$ hours. It should be noted that this generation of cells is very difficult to distinguish in this particular histogram data set. Such poorly resolved generations of cells can be quite problematic for the deconvolution techniques, as the unique estimation of parameters (for the normal or lognormal curves) requires that distinct generations of cells be plainly visible. It appears to be a major advantage of the compartmental model to be able to fit data (and hence compute cell numbers) even when the histogram data features generations of cells which are less than ideally resolved. Of course, it is not possible to say from these results which technique (if either) is providing the correct number of cells. Yet, because the compartmental model is derived from a conservation law, and this conservation law must hold regardless of the parameters input into the model, cells cannot enter or leave the population except as permitted by the form of the model and the given parameters. Meanwhile, the deconvolution techniques do not arise from any conservation law, and the computed cell numbers in each generation may increase or decrease freely, unrestrained by any balance law. It seems then, that the

Table 8: Total precursors in terms of divisions number. For each time and generation, the total number of precursors has been computed from the OLS best-fit model solution (top), from a deconvolution of the data using normal curves (middle) and from a deconvolution of the data using lognormal curves (bottom). As in Table 7 we find general agreement between the values computed from deconvolution techniques, which are slightly different than those computed from the compartmental model. Under the assumptions of the experiment, the total number of precursors should not increase.

	Divisions Undergone							
Time (hrs)	0	1	2	3	4	5	6	Total
24	11339892	0	0	0	0	0	0	11339892
	9211571	0	0	0	0	0	0	9211571
	9254681	0	0	0	0	0	0	9254681
48	5881557	3277907	0	0	0	0	0	9159465
	6298359	2736564	0	0	0	0	0	9034923
	6294945	2717285	0	0	0	0	0	9012230
96	1906065	1239021	2023141	2622054	1157526	237156	0	9184963
	1970401	1642000	2754838	2273013	1144224	164136	0	9948612
	2364520	1970400	2616943	2221689	1177916	168969	0	10520436
120	1476266	989465	1401482	2135859	1595582	812070	224017	8634740
	1969773	1484648	1970400	2627200	1642000	779950	71837	10545808
	2195762	1717837	1930663	2143761	1672236	935940	86205	10682405

compartmental model should have a major advantage in computing cell numbers, owing to its 'memory' of the number of cells determined at previous time points (even when the generations of cells are poorly resolved in the data).

This is particularly noteworthy in Table 8, where the number of precursors for each generation of cells is computed. As in Table 7, each block of Table 8 contains the results computed from the compartmental model, deconvolution with normal curves, and deconvolution with lognormal curves. Observe the significant increase (more than 10%) in the total number of precursors as computed by deconvolution between $t = 48$ and $t = 96$ hours. As discussed above, the total number of precursors cannot increase in a population of cells. While the number of precursors computed from the compartmental model also increases, it does so by a very small amount (less than 0.3%) consistent with the error in the numerical solver. Of course, it has already been noted that some data sets do in fact exhibit increases in the total number of precursors–a discrepancy arising from the inaccuracy of the assumption that each well plate contains an identical population of cells. On one hand, the deconvolution techniques would seem to have an advantage, as they are not constrained by any conservation law. However, this has an interesting implication. Because the deconvolution techniques do not link the population estimates from one data collection time to the next, there is a potential bias associated with such methods as a result of sample-to-sample variability in the experimental data. It should be noted that sample-to-sample variability is also problematic for the compartmental model solution. If the samples used to obtain the experimental data are not sufficiently similar, the conservation law (which follows from the assumption that each sample is identical) used to derive the model may not hold. In such a case, the compartmental model would be systematically in error (when compared to the data), as the calibrated model itself would still follow the assumed conservation law. Following the discussion above, we believe that a more accurate statistical model, which will necessarily include a careful consideration of the method of sampling/data collection, will resolve any discrepancy with the compartmental model. Some preliminary work on this subject is surveyed in [66, Ch. 4].

5.3 Generalizations of the Mathematical Model

Apart from issues involving the statistical model relating the mathematical model to the data, it has been shown that the compartmental model accurately reproduces the behavior of a PHA-stimulated population of CD4+ cells as represented in histogram data from a flow cytometry assay. This model accounts for the natural rate of CFSE FI decay resulting from turnover of the intracellular label as well as the autofluorescence of cells in the absence

of any fluorescent labeling. Simple linear models are used to describe the rates of cell division and death.

At the moment, only a single CFSE data set has been examined and used to estimate the parameters of the mathematical model(s). It is believed that the compartmental model is quite general and should apply to a wide range of data sets from various experimental setups. Work is ongoing to collect additional data sets to demonstrate such a wide applicability of the model. As additional data sets become available, several additional features may need to be considered at greater length.

It is hoped that the compartmental model can be generalized to account for multiple cell types both in vivo and in vitro. While the cells studied in this report were cultured in a saturating quantity of the stimulating agent PHA, cells in vivo (or even cells in vitro in a different experimental setup) will not experience such a strong, constant stimulation. As such, the possibility exists that some cells may return to a quiescent state during the proliferation assay. It is known that the autofluorescence of a cell changes depending upon its state of activation, and thus this mechanism may need to be included in subsequent modeling efforts. (For the current data set, the quiescent cells are all undivided, and AutoFI is negligible for those cells.)

Similar to the efforts in [18, 19], we have used Malthusian rates for both proliferation (with time-dependent rates $\alpha_i(t)$) and death (with rates β_i). As discussed in Section 3, such an assumption is reasonable provided the turnover of cells (resulting either from division or death) occurs at a sufficiently rapid pace. Given the physiological constraints placed on rapidly dividing cells (e.g., rates of growth and DNA replication), one would expect some sort of minimum cell cycle time. It is unclear if the necessity of time dependence in the Malthusian rates α_i is an artifact of such a feature. To test this hypothesis, several generalizations of the proliferation and death rate terms are immediately available.

First, one might consider the addition of a second structure variable (say, volume) which could be used to enforce a minimum cell cycle time by requiring that cells progress from some size V to $2V$ before dividing, at which point two cells of size V are produced. However, in the absence of additional observations, it is unclear what parameters (e.g., average rate of growth, or the parameter V) might be estimable from CFSE histogram data. Video microscopy measurements by Hawkins, et al. [41] indicate that average cell size may be division dependent, and this may add some additional complexity to the inclusion of volume structure. Biologically, it is expected that apoptosis occurs only at particular checkpoints in the cell cycle (particularly if external 'kill signals' are absent) so that a generalization to volume structure (or any other surrogate for cell cycle position or physiological age) may permit a more accurate description of cell death. Still, it is unclear what information might be estimated from only CFSE histogram data. It is possible that the forward scatter (FSC) of laser light may possibly be used as an observable surrogate for cell size, and some additional work will be necessary to investigate this possibility.

A second possibility to generalize the rates of proliferation and death would be to consider rate-limiting (e.g., logistic, Gompertz) models for proliferation and death. Some biological mechanisms have been proposed which may lead to density-dependent rates of cell death [26], and a Gompertz model for cell growth has been used to account for quiescence in the context of a size-structured population model [39]. Of course, generalizations to nonlinear division and death must be considered in the context of the improvement they provide in fitting a given model to CFSE data sets. Given the accuracy of the simple linear models (albeit with time-dependent rates of proliferation), such generalizations seem unnecessary at the moment.

Given that the compartmental model can be used to compute numbers of cells per generation directly, some comparison has already been made between the results obtained with this model and the cell numbers computed from deconvolution techniques (Tables 7 and 8). It remains to compare the parameter estimates and model fits obtained with the compartmental model with those obtained from previous models (Smith-Martin, cyton, etc.). In fact, it maybe possible to incorporate into the compartmental model the mathematical forms used to describe proliferation and death in these models. Recall that the method of characteristics provides a solution (Equations 9 and 11) of the form

$$n_i(t, x(t; s)) = F(\text{Division,Death}) \tag{34}$$

where $x(t; s)$ is the characteristic line emanating from the point $(0, s)$ in the tx-plane. Clearly, the left side of Equation (34) is independent of any mathematical formulation of cell proliferation and death. In this report, the form of the hypothetical function F is determined from the PDE formulation of the compartmental model (2) and the accompanying assumptions regarding the Malthusian rates of proliferation and death. Alternative, one could consider using (34) or its differential form (i.e., (10)) as a starting point, defining the right side of

the equation in accordance with the assumptions of the Smith-Martin or cyton models, or their generalizations [33, 40, 46, 57, 71]. While previous authors have derived these models specifically in terms of total cell numbers, (34) could be related back to previous work by simple integration. The primary advantage in using (34) would be in the direct comparison of the model to histogram data, rather than from computed cell numbers. Further study could reveal the extent (if any) to which such a direct comparison improves the unique identification of parameters in previous models, although this will first rely on an accurate statistical model.

In this context, it is clear that several alternative possibilities exist for a mathematical description of proliferation and death rates. Thus it is clear that the interpretation of the proliferation and death parameters must be made with careful regard to the form of the model. Given the form of the model solution (Equations (9) and (11)) for the compartmental model, it is plainly observed that linear changes in parameters for proliferation and death rates cause an exponential response in the computed solution [37, 31]. As such, the sensitivity of the model to these parameters, as well as the degree to which their estimation is unique, must be carefully considered when interpreting estimated parameters. The uniqueness of the estimated functions $\alpha_i(t)$ will depend on how the nodes for the linear splines are chosen in relation to the times at which data is taken. In some models, it has been shown that the effects of a linear increase of cell cycle time with division number cannot be distinguished from the effects of a linear increase in the death rate with division number [47]. If this is the case, then the biological interpretation of some parameters may be suspect.

Ideally, the values of $\alpha_i(t)$ and β_i can be related back to more physical/experimental parameters such as the type and strength of stimulation, which may in turn require the mathematical modeling of certain molecular pathways within individual cells. Recent work has indicated that the mechanisms responsible for cell proliferation and death may be mutually dependent upon a common molecular pathway [32, 65]. As more data becomes available, we hope to examine how the estimated parameters change under various experimental conditions, with an eye toward additional constitutive relationships linking molecular and/or subcellular functions to population dynamics [24]. In this context, it seems necessary to consider the extent to which these functions and/or pathways are inherited. Evidence suggests that closely related cells exhibit strong correlation in times to divide and some correlation in times to die, and that this correlation tends to decrease with the number of divisions undergone [41]. Cells with a common precursor may also share a common division destiny [41], which can be altered by stimulation conditions [68]. While computed cell numbers are relatively unaffected provided correlation is limited to cells having undergone the same number of divisions [33, 41, 44], correlation between subsequent division of cells can alter the dynamics predicted by a mathematical model [71]. For large populations, this effect seems negligible, but may play an important role in vivo where only a small number of responding cells can trigger an immune response [71]. Cyton models and branching process models have been formulated to account for various levels of correlation [33, 44, 71], and these models may be incorporated into the compartmental model framework as described above. Alternatively, it may be possible (given any reasonable, identifiable parameterizations of cell division and death) to place probability distributions on these parameters (e.g., on the functions $\alpha_i(t)$ and $\beta_i(t)$) [6, 9, 12] in the manner described in Section 3.3.

5.4 Concluding Remarks

The compartmental model is the latest in a series of structured PDE models which can be fit directly to histogram representations of flow cytometry data. Once calibrated, the compartmental model can be used to quickly and accurately estimate the numbers of cells having undergone a certain number of divisions. This information can be used to determine biologically relevant parameters which will help to meaningfully compare cells from different donors and experiments. While the use of cell numbers per generation is not new, the direct modeling of histogram data reduces any need for deconvolution techniques which may introduce unnecessary bias into the computed cell numbers. Moreover, because the model is based upon conservation principles, it should be possible to fit histogram data even when the 'peaks' in the data (representing distinct generations of cells) are not well-resolved. This is a significant advantage over deconvolution techniques. The actual number of generations which can be accurately modeled (that is, the maximum value of $i_{\max}$) will depend upon the uniformity of the initial uptake of intracellular dye as well as the magnitude of the resulting CFSE FI relative to cellular AutoFI.

We are actively working to collect additional data sets with which to demonstrate the widespread applicability of this model, as well as to use this model in a systematic fashion to analyze how the estimated parameters

vary under changing experimental and biological conditions. Most immediately, this will require the development of an accurate statistical model for the data. The generalization of the model to multiple cell types is immediate, although an accurate quantification of any interaction terms will require some careful thought and experimentation.

As more information becomes available regarding the complex processes involved in cell proliferation, we are confident that the model discussed here provides a firm physiological foundation upon which CFSE-based assay data can be understood. We strongly believe that the ideas and results presented here will form an important interpretive framework with a wide array of applications in experimental settings, diagnostic tests [34], and perhaps in a more integrated model of cell dynamics [45, 49].

Acknowledgments:

This research was supported in part by the National Institute of Allergy and Infectious Disease under grant NIAID 9R01AI071915, in part by the U.S. Air Force Office of Scientific Research under grant AFOSR-FA9550-09-1-0226, in part by the Deutsche Forschungsgemeinschaft and in part by grant SAF2010-21336 from the Spanish Ministry of Science and Innovation.

References

[1] H.T. Banks, *A Functional Analysis Framework for Modeling, Estimation and Control in Science and Engineering*, CRC Press/Taylor-Francis, Boca Raton London New York, 2012.

[2] H.T. Banks and Kathleen Bihari, Modelling and estimating uncertainty in parameter estimation, *Inverse Problems*, **17** (2001), 95–111.

[3] H. T.Banks, V. A. Bokil, S. Hu, F. C. T. Allnutt, R. Bullis, A. K. Dhar and C. L. Browdy, Shrimp biomass and viral infection for production of biological countermeasures, CRSC-TR05-45, December, 2005; *Mathematical Biosciences and Engineering*, **3** (2006), 635–660.

[4] H. T. Banks, D. M. Bortz and S. E. Holte, Incorporation of variability into the mathematical modeling of viral delays in HIV infection dynamics, *Math. Biosciences*, **183** (2003), 63–91.

[5] H. T. Banks, D. M. Bortz, G. A. Pinter and L. K. Potter, Modeling and imaging techniques with potential for application in bioterrorism, CRSC-TR03-02, January, 2003; Chapter 6 in *Bioterrorism: Mathematical Modeling Applications in Homeland Security*, (H.T. Banks and C. Castillo-Chavez, eds.), Frontiers in Applied Math, **FR28**, SIAM, Philadelphia, PA, 2003, 129–154.

[6] H.T. Banks, L.W. Botsford, F. Kappel, and C. Wang, Modeling and estimation in size structured population models, LCDS/CSS Report 87-13, Brown University, March 1987; *Proc. 2nd Course on Math. Ecology*, (Trieste, December 8-12, 1986) World Scientific Press, Singapore, 1988, 521–541.

[7] H.T. Banks, Frederique Charles, Marie Doumic, Karyn L. Sutton, and W. Clayton Thompson, Label structured cell proliferation models, *Appl. Math. Letters*, **23** (2010), 1412–1415; doi:10.1016/j.aml.2010.07.009

[8] H.T. Banks, M. Davidian, J. Samuels, and K.L. Sutton, An inverse problem statistical methodology summary, CRSC-TR08-01, NCSU, January, 2008; Chapter 11 in *Mathematical and Statistical Estimation Approaches in Epidemiology*, G. Chowell, et al., eds., Berlin Heidelberg New York, 2009, 249–302.

[9] H.T. Banks and J.L. Davis, A comparison of approximation methods for the estimation of probability distributions on parameters, *Appl. Num. Math.*, **57** (2007), 753–777.

[10] H. T. Banks, J. L. Davis, S. L. Ernstberger, S. Hu, E. Artimovich, A. K. Dhar and C. L. Browdy, A comparison of probabilistic and stochastic formulations in modeling growth uncertainty and variability, CRSC-TR08-03, February, 2008; *Journal of Biological Dynamics*, **3** (2009), 130–148.

[11] H.T. Banks and B.G. Fitzpatrick, Inverse problems for distributed systems: statistical tests and ANOVA, LCDS/CSS Report 88-16, Brown University, July 1988; *Proc. International Symposium on Math. Approaches to Envir. and Ecol. Problems*, Springer Lecture Notes in Biomath., **81** (1989), 262–273.

[12] H.T. Banks and B.F. Fitzpatrick, Estimation of growth rate distributions in size-structured population models, CAMS Tech. Rep. 90-2, Univ. of Southern California, January, 1990; *Quart. Appl. Math.* **49** (1991), 215–235.

[13] H. T. Banks and N. L. Gibson, Well-posedness in Maxwell systems with distributions of polarization relaxation parameters, CRSC-TR04-01, January, 2004; *Applied Math. Letters*, **18** (2005), 423–430.

[14] H. T. Banks and N. L. Gibson, Electromagnetic inverse problems involving distributions of dielectric mechanisms and parameters, CRSC-TR05-29, August, 2005; *Quarterly of Applied Mathematics*, **64** (2006), 749–795.

[15] H.T. Banks and K. Kunisch, *Estimation Techniques for Distributed Parameter Systems*, Birkhauser, Boston, 1989.

[16] H. T. Banks and G. A. Pinter, A probabilistic multiscale approach to hysteresis in shear wave propagation in biotissue, CRSC-TR04-03, January, 2004; *SIAM J. Multiscale Modeling and Simulation*, **3** (2005), 395–412.

[17] H. T. Banks and L. K. Potter, Probabilistic methods for addressing uncertainty and variability in biological models: Application to a toxicokinetic model, CRSC-TR02-27, September, 2002; *Math. Biosci.*, **192** (2004), 193–225.

[18] H.T. Banks, Karyn L. Sutton, W. Clayton Thompson, G. Bocharov, Marie Doumic, Tim Schenkel, Jordi Argilaguet, Sandra Giest, Cristina Peligero, and Andreas Meyerhans, A New Model for the Estimation of Cell Proliferation Dynamics Using CFSE Data, CRSC-TR11-05, NCSU, Revised July, 2011; *J. Immunological Methods,* **373** (2011), 143–160; doi:10.1016/j.jim.2011.08.014.

[19] H.T. Banks, Karyn L. Sutton, W. Clayton Thompson, Gennady Bocharov, Dirk Roose, Tim Schenkel, and Andreas Meyerhans, Estimation of cell proliferation dynamics using CFSE data, CRSC-TR09-17, NCSU, August, 2009; *Bull. Math. Biol.* **70** (2011), 116–150; doi:10.1007/s11538-010-9524-5.

[20] H.T. Banks and H.T. Tran, *Mathematical and Experimental Modeling of Physical and Biological Processes*, CRC Press, Boca Raton London New York, 2009.

[21] H.T. Banks, B.G. Fitzpatrick, Laura K. Potter, and Yue Zhang, Estimation of probability distributions for individual parameters using aggregate population observations, CRSC-TR98-06, January, 1998; *Stochastic Analysis, Control, Optimization and Applications* (W.McEneaney, G. Yin, and Q. Zhang, eds.), Birkhäuser, (1998), 353-371.

[22] G. Bell and E. Anderson, Cell growth and division I. A mathematical model with applications to cell volume distributions in mammalian suspension cultures, *Biophysical Journal*, **7** (1967), 329–351.

[23] K.P. Burnham and D.R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, (2nd Edition), Springer, New York, 2002.

[24] Nigel J. Burroughs and P. Anton van der Merwe, Stochasticity and spatial heterogeneity in T-cell activation, *Immunological Reviews*, **216** (2007), 69–80.

[25] R. Callard and P.D. Hodgkin, Modeling T- and B-cell growth and differentiation, *Immunological Reviews*, **216** (2007), 119–129.

[26] Robin E. Callard, Jaroslav Stark, and Andrew J. Yates, Fratricide: a mechanism for T memory-cell homeostasis, *Trends in Immunology*, **24** (2003), 370–375.

[27] R.J. Carroll and D. Ruppert, *Transformation and Weighting in Regression*, Chapman Hall, London, 2000.

[28] M. Davidian and D.M. Giltinan, *Nonlinear Models for Repeated Measurement Data*, Chapman and Hall, London, 2000.

[29] R.J. DeBoer, V.V. Ganusov, D. Milutinovic, P.D. Hodgkin, and A.S. Perelson, Estimaing lymphocyte division and death rates from CFSE data, *Bull. Math. Biol.*, **68** (2006), 1011–1031.

[30] R.J. DeBoer and Alan S. Perelson, Estimating division and death rates from CFSE data, *J. Comp. and Appl. Mathematics*, **184** (2005), 140–164.

[31] E.K. Deenick, A.V. Gett, P.D. Hodgkin, Stochastic model of T cell proliferation: a calculus revealing IL-2 regulation of precursor frequencies, cell cycle time, and survival, *J. Immunology*, **170** (2003), 4963–4972.

[32] Mark R Dowling, Dejan Milutinovic, and Philip D Hodgkin, Modelling cell lifespan and proliferation: is likelihood to die or to divide independent of age?, *J. R. Soc. Interface*, **2** (2005), 517–526.

[33] K. Duffy and V. Subramanian, On the impact of correlation between collaterally consanguineous cells on lymphocyte population dynamics, *J. Math. Biol.*, **59** (2009), 255–285.

[34] D.A. Fulcher and S.W.J. Wong, Carboxyfluorescein diacetate succinimidyl ester-based assays for assessment of T cell function in the diagnostic laboratory, *Immunology and Cell Biology*, **77** (1999), 559–564.

[35] Vitaly V. Ganusov, Dejan Milutinovi, and Rob J. De Boer, IL-2 regulates expansion of CD4+ T cell populations by affecting cell death: insights from modeling CFSE data, *J. Immunology*, **179** (2007), 950–957.

[36] V.V. Ganusov, S.S. Pilyugin, R.J. De Boer, K. Murali-Krishna, R. Ahmed, and R. Antia, Quantifying cell turnover using CFSE data, *J. Immunological Methods*, **298** (2005), 183–200.

[37] A.V. Gett and P.D. Hodgkin, A cellular calculus for signal integration by T cells, *Nature Immunology*, **1** (2000), 239–244.

[38] M. Kot, *Elements of Mathematical Ecology*, Cambridge University Press, Cambridge, UK, 2001.

[39] M. Gyllenberg and G. F. Webb, A nonlinear structured population model of tumor growth with quiescence, *J. Math. Biol.*, **28** (1990), 671–694.

[40] E.D. Hawkins, Mirja Hommel, M.L Turner, Francis Battye, J Markham and P.D Hodgkin, Measuring lymphocyte proliferation, survival and differentiation using CFSE time-series data, *Nature Protocols*, **2** (2007), 2057–2067.

[41] E.D. Hawkins, J.F. Markham, L.P. McGuinness, and P.D. Hodgkin, A single-cell pedigree analysis of alternative stochastic lymphocyte fates, *Proc. Natl. Acad. Sci.*, **106** (2009), 13457–13462.

[42] Mirja Hommel and Philip D. Hodgkin, TCR affinity promotes CD8+ T-cell expansion by regulating survival, *J. Immunology*, **179** (2007), 2250–2260.

[43] O. Hyrien and M.S. Zand, A mixture model with dependent observations for the analysis of CFSE-labeling experiments, *J. American Statistical Association*, **103** (2008), 222–239.

[44] O. Hyrien, R. Chen, and M.S. Zand, An age-dependent branching process model for the analysis of CFSE-labeling experiments, *Biology Direct*, **5** (2010), Published Online.

[45] D.E. Kirschner, S.T. Chang, T.W. Riggs, N. Perry, and J.J. Linderman, Toward a multiscale model of antigen presentation in immunity, *Immunological Reviews*, **216** (2007), 93–118.

[46] H.Y. Lee, E.D. Hawkins, M.S. Zand, T. Mosmann, H. Wu, P.D. Hodgkin, and A.S. Perelson, Interpreting CFSE obtained division histories of B cells in vitro with Smith-Martin and Cyton type models, *Bull. Math. Biol.*, **71** (2009), 1649–1670.

[47] H.Y. Lee and A.S. Perelson, Modeling T cell proliferation and death in vitro based on labeling data: generalizations of the Smith-Martin cell cycle model, *Bull. Math. Biol.*, **70** (2008), 21–44.

[48] K. Leon, J. Faro, and J. Carneiro, A general mathematical framework to model generation structure in a population of asynchronously dividing cells, *J. Theoretical Biology*, **229** (2004), 455–476.

[49] Y. Louzoun, The evolution of mathematical immunology, *Immunological Reviews*, **216** (2007), 9–20.

[50] T. Luzyanina, D. Roose, and G. Bocharov, Distributed parameter identification for a label-structured cell population dynamics model using CFSE histogram time-series data, *J. Math. Biol.*, **59** (2009), 581–603.

[51] T. Luzyanina, M. Mrusek, J.T. Edwards, D. Roose, S. Ehl, and G. Bocharov, Computational analysis of CFSE proliferation assay, *J. Math. Biol.*, **54** (2007), 57–89.

[52] T. Luzyanina, D. Roose, T. Schenkel, M. Sester, S. Ehl, A. Meyerhans, and G. Bocharov, Numerical modelling of label-structured cell population growth using CFSE distribution data, *Theoretical Biology and Medical Modelling*, **4** (2007), Published Online.

[53] A.B. Lyons and C.R. Parish, Determination of lymphocyte division by flow cytometry, *J. Immunol. Methods*, **171** (1994), 131–137.

[54] A. B. Lyons, J. Hasbold and P.D. Hodgkin, Flow cytometric analysis of cell division history using diluation of carboxyfluorescein diacetate succinimidyl ester, a stably integrated fluorescent probe, *Methods in Cell Biology*, **63** (2001), 375–398.

[55] G. Matera, M. Lupi and P. Ubezio, Heterogeneous cell response to topotecan in a CFSE-based proliferative test, *Cytometry A*, **62** (2004), 118–128.

[56] J.A. Metz and O. Diekmann, *The Dynamics of Physiologically Structured Populations*, Springer Lecture Notes in Biomathematics, **68** (1986).

[57] Robert E. Nordon, Kap-Hyoun Ko, Ross Odell, and Timm Schroeder, Multi-type branching models to describe cell differentiation programs, *J. Theoretical Biology*, **277** (2011), 7–18.

[58] R.E. Nordon, M. Nakamura, C. Ramirez, and R. Odell, Analysis of growth kinetics by division tracking, *Immunology and Cell Biology*, **77** (1999), 523–529.

[59] C. Parish, Fluorescent dyes for lymphocyte migration and proliferation studies, *Immunology and Cell Biol.*, **77** (1999), 499–508.

[60] B. Quah, H. Warren and C. Parish, Monitoring lymphocyte proliferation in vitro and in vivo with the intracellular fluorescent dye carboxyfluorescein diacetate succinimidyl ester, *Nature Protocols*, **2** (2007), 2049–2056.

[61] P. Revy, M. Sospedra, B. Barbour, and A. Trautmann, Functional antigen-independent synapses formed between T cells and dendritic cells, *Nature Immunology*, **2** (2001), 925–931.

[62] G.A. Sever and C.J. Wild, *Nonlinear Regression*, Wiley, Hoboken, NJ, 2003.

[63] J. Sinko and W. Streifer, A new model for age-size structure of a population, *Ecology*, **48** (1967), 910–918.

[64] V.G. Subramanian, K.R. Duffy, M.L. Turner and P.D. Hodgkin, Determining the expected variability of immune responses using the cyton model, *J. Math. Biol.*, **56** (2008), 861–892.

[65] David T. Terrano, Meenakshi Upreti and Timothy C. Chambers, Cyclin-dependent kinase 1-mediated Bcl-x_L/Bcl-2 phosphorylation acts as a functional link coupling mitotic arrest and apoptosis, *Mol. Cell. Biol.*, **30** (2010), 640–656.

[66] W. Clayton Thompson, *Partial Differential Equation Modeling of Flow Cytometry Data from CFSE-based Proliferation Assays*, Ph.D. Dissertation, North Carolina State University, December, 2011.

[67] B. Tummers, DataThief III. 2006 (http://datathief.org/)

[68] M.L. Turner, E.D. Hawkins, and P.D. Hodgkin, Quantitative regulation of B cell division destiny by signal strength, *J. Immunology*, **181** (2008), 374–382.

[69] H. Veiga-Fernandez, U. Walter, C. Bourgeois, A. McLean, and B. Rocha, Response of naive and memory CD8+ T cells to antigen stimulation in vivo, *Nature Immunology*, **1** (2000), 47–53.

[70] P.K. Wallace, J.D. Tario, Jr., J.L. Fisher, S.S. Wallace, M.S. Ernstoff, and K.A. Muirhead, Tracking antigen-driven responses by flow cytometry: monitoring proliferation by dye dilution, *Cytometry A*, **73** (2008), 1019–1034.

[71] C. Wellard, J. Markham, E.D. Hawkins, and P.D. Hodgkin, The effect of correlations on the population dynamics of lymphocytes, *J. Theoretical Biology*, **264** (2010), 443–449.

[72] J.M. Witkowski, Advanced application of CFSE for cellular tracking, *Current Protocols in Cytometry*, (2008), 9.25.1–9.25.8.

[73] A. Yates, C. Chan, J. Strid, S. Moon, R. Callard, A.J.T. George, and J. Stark, Reconstruction of cell population dynamics using CFSE, *BMC Bioinformatics*, **8** (2007), Published Online.